ESCUELA POLITÉCNICA NACIONAL

SECOND-ORDER DESCENT METHODS AND ACTIVE-SET STRATEGIES FOR NON-SMOOTH OPTIMIZATION WITH APPLICATIONS TO VISCOPLACTIC FLUIDS AND GROUP SPARSE OPTIMIZATION

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE DOCTORADO EN MATEMÁTICA APLICADA

TESIS

SOFÍA ALEJANDRA LÓPEZ ORDÓÑEZ
sofia.lopezo@epn.edu.ec

Director: SERGIO ALEJANDRO GONZÁLEZ ANDRADE PHD
sergio.gonzalez@epn.edu.ec

Codirector: PEDRO MARTÍN MERINO ROSERO PHD
pedro.merino@epn.edu.ec

QUITO, ENERO 2025

DECLARACIÓN

Yo, SOFÍA ALEJANDRA LÓPEZ ORDÓÑEZ declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Sofía Alejandra López Ordóñez

CERTIFICACIÓN

We certify that the present work was developed by SOFÍA ALEJANDRA LÓPEZ ORDÓÑEZ, under our supervision.

Sergio Alejandro González Andrade PhD
Director de Tesis

Pedro Martín Merino Rosero PhD
Codirector de Tesis

ACKNOWLEDGMENTS

This project would not have been possible without the support of many people. I want to thank my advisors, Sergio González Andrade and Pedro Merino for their patience and guidance throughout this challenging journey. Their valuable feedback and thoughful conversations motivated me to sharpen my thinking. Thank you for your ideas and advice throughout these years.

I am also thankful to Juan Carlos De los Reyes for his guidance and very helpful discussions on my research topic. His insightful feedback provided me with the tools to improve my understandings and find the path for my dissertation.

I also want to thank my colleagues at the Research Center on Mathematical Modeling - Modemat for their wonderful collaboration in these years. I want to thank Luis Miguel Torres and Diego Recalde, for their support and invaluable help.

I want to thank David Villacís, Evelyn Cueva, Santiago Núñez, Diana Taipe, Irwin Jiménez, Edison Vera, Myrian Guanoluiza, Felipe Fernández, María Belén Moya, Kateryn Herrera, Paula Castro and Mercy Anchundia for their invaluable support and friendship.

Finally, I want to deeply thank my family for their encouragement.

This thesis was developed under the financial support of Secretaría de Educación Superior, Ciencia, Tecnología e Innovación - Senescyt and the project EPN-PIGR 19-02.

To Enrique, Bibiana, Daniela, David and Joaquín.

Contents

List of Algorithms

Abstract

This thesis focuses on the development and analysis of numerical optimization methods for solving nonsmooth problems arising in viscoplastic fluid dynamics and group-sparse regularization problems. The primary application centers on the Bingham flow problem, a type of yield-stress fluid characterized by a transition between solid-like and fluid-like behavior depending on a stress threshold. To address the incompressibility condition and nonsmooth terms inherent in the Bingham fluid problem, we propose a second-order descent algorithm that incorporates exact penalization, generalized second-order information, and active-set strategies.

In the first part of this thesis, we analyze the exact penalization of the incompressibility constraint using the L^1-norm penalization within a regularized formulation of the Bingham flow problem. We establish a penalization parameter that ensures equivalence between the penalized and constrained formulations and develop an algorithm that uses second-order information to solve the resulting nonsmooth optimization problem efficiently. Numerical experiments confirm the effectiveness of the approach, particularly in promoting sparsity on the divergence term (incompressibility constraint term), and comparing it with the Semi-smooth Newton method.

The second part of this thesis is motivated by the solution of the unregularized Bingham flow problem. A simplified version of this problem is reformulated as a linearly constrained minimization problem, where the constrained term, interpreted as a group-sparsity regularizer, identifies regions where the material behaves like a rigid solid. The problem is analyzed through its augmented Lagrangian formulation, and a specialized group-sparse algorithm is proposed to solve it.

Building on the insights gained from the unregularized Bingham flow problem, these strategies are extended to a more general group-sparse optimization framework involving the $\|\cdot\|_{1,2}$ norm, with applications to PDE-constrained optimization and nonlinear regression problems. A second-order algorithm is derived, relying on the computation of the steepest descent direction, which is determined through group-wise evaluation of the directional derivative.

Additionally, an active-set identification strategy is introduced. The novelty of this approach lies in its dynamic and efficient identification of sparse groups by iteratively

analyzing the angle between two consecutive iterations. This active-set strategy is further combined with a reduced second-order system to enhance computational efficiency and improve accuracy, making the proposed algorithms effective for addressing nonsmooth optimization problems.

In summary, the second part of the thesis focuses on the development of group-sparse algorithms with applications to the unregularized Bingham flow problem in a pipe and group-sparse optimization problems.

The algorithms developed in this thesis are analyzed with extensive numerical validation which highlights their effectiveness.

Chapter 1

Introduction

1.1 Motivation

Viscoplastic fluids are materials that behave like solids at low stress levels but flow like liquids when the applied stress exceeds a certain threshold, known as the yield stress [70]. Below this threshold, they resist deformation and do not flow. Once the yield stress is surpassed, they begin to deform and flow, displaying a combination of viscous and plastic behavior. Common examples of viscoplastic fluids include toothpaste, ketchup, and some drilling muds. However, while toothpaste is a classic example of a Bingham fluid, the implications of modeling such flows extend far beyond this illustrative case. Bingham flow models are relevant in many industrial and geophysical contexts. Fresh concrete and certain slurries used in mining and drilling are modeled as Bingham fluids [41]. Predicting their flow is key for structural casting and avoiding clogging or segregation in transport systems. Certain biological fluids, such as blood under specific conditions (e.g., in microcirculation or in clotting scenarios), may display Bingham-like behavior [116]. Moreover, lava flows are often modeled using Bingham-type fluids due to their yield stress and complex rheology (see [20] and [73]). Thus, reliable numerical simulations aid in hazard assessment and risk mitigation in volcanic regions. Particularly, lava flow is simulated by viscoplastic flow in a shallow-water regime using the Bingham model. In [20], researchers use the augmented Lagrangian method with well-balanced properties coupled with the finite volume discretization to solve one-dimensional Bingham viscoplastic flow. A key challenge here is coupling the shallow-water equations with the viscoplastic constitutive laws while addressing high computational costs. Thus, numerical modeling of these fluids is important for design, safety, and operational efficiency. The theoretical and numerical developments presented in this thesis may be extended to the shallow-water framework for modeling lava flows, providing a potential pathway for simulating viscoplastic behavior in this setting.

Numerical flow simulation for viscoplastic fluids is an active research field which faces challenges related to the nonlinear and nonsmooth nature of these viscoplastic materials. Mathematically, the fluid's velocity field can be characterized through a variational inequality, where the minimizer of the inequality serves as the solution to the flow problem [46]. Thus, variational inequalities and nonsmooth optimization techniques are directly applicable for an accurate modeling of a viscoplastic flow.

The divergence-free condition, often expressed as $\operatorname{div}(\mathbf{u}) = \mathbf{0}$, where $\mathbf{u}$ is the velocity field of the fluid, is a critical aspect in the numerical simulation of viscoplastic fluids for several reasons: to ensure the incompressibility of the fluids in simulations by coupling the velocity and pressure fields to preserve the physical accuracy of the model and ensure realistic and stable results. In particular, the rheology of viscoplastic fluids is highly sensitive to stress and strain rates. Therefore, the divergence-free condition becomes even more critical to accurately capture the flow dynamics and transition between yielded and unyielded regions. Failing to enforce the divergence-free condition can lead to non-physical results such as artificial compressibility or spurious oscillations in the velocity and pressure fields.

Enforcing the divergence-free condition introduces a distinct challenge to address. Finite Element Methods such as the $H(\operatorname{div})$-conforming methods [27, 63, 103] are used for flow problems to address the incompressibility condition. Nonetheless, mesh refinement techniques or the use of high-order polynomials play a significant role in enhancing the accuracy and stability of these simulations [75, 82]. This motivates addressing the divergence-free constraint from an optimization perspective. Consequently, the use of penalization techniques for the divergence-free condition is the first main focus of this thesis.

The close connection between nonsmooth optimization problems and flow problems in viscoplasticity arises from the yield stress characterization of viscoplastic fluids. The transition between solid-like and fluid-like regimes creates a nonsmooth response in the stress-strain relationship, resulting in a nonsmooth optimization problem. Approaching the flow problem as a nonsmooth problem requires mathematical tools and methods from nonsmooth analysis. For instance, traditional first-order algorithms for numerically solving the convex nonsmooth optimization flow problem are derived from the augmented Lagrangian methods [42, 49, 71], see also [112]. Thus, the second main focus of this thesis is to derive a second-order optimization algorithm for the solution of the resulting non-smooth problem arising from the viscoplastic fluid flow. Additionally, the nonsmooth term arising in the viscoplastic flow problem is defined in terms of the norm of the strain-rate tensor, which describes the rate of change of the relative deformation of the material. This tensor is null where the material behaves like a rigid solid. Consequently, in the Bingham flow problem, this nonsmooth term can be interpreted as a sparsity-promoting term in the solid-like regime. Building on this

perspective, we will extend the concepts studied for this term to more general sparsity structures such as the group-sparse term given by the $\|\cdot\|_{1,2}$ norm.

1.2 Introduction

In the first part of this thesis, we are devoted to the analysis and design of a descent optimization algorithm that incorporates second-order information to solve a viscoplastic fluid problem. Specifically, our attention is directed towards Bingham fluids, which represent a type of yield stress fluid.

The flow of a Bingham fluid can be described in terms of a variational inequality of the second kind [46, Ch. VI]. This variational inequality constitutes the optimality condition of the following nondifferentiable convex optimization problem:

$$\begin{cases} \min\limits_{\mathbf{u}\in\mathbf{H}_0^1(\Omega)} \widetilde{J}(\mathbf{u}) := \mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{u}\, dx + g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx - \int_\Omega \mathbf{f_b}\cdot\mathbf{u}\, dx \\ \text{s.t} \quad \operatorname{div}\mathbf{u} = 0. \end{cases} \tag{1.1}$$

Its solution corresponds to the velocity field $\mathbf{u}$ of the steady-state Bingham flow. We consider Ω an open and bounded subset of $\mathbb{R}^n$, for $n = 2, 3$, with Lipschitz boundary, $\mu > 0$ corresponds to the viscosity parameter, $\operatorname{div} : \mathbf{H}_0^1(\Omega) \to L_0^2(\Omega)$ represents the divergence operator, $\mathcal{E} = \frac{1}{2}(\nabla + \nabla^\top)$ is the symmetric gradient operator (strain rate tensor), and $\mathbf{f_b}$ are the body forces acting on the fluid or a pressure gradient. The notation $:$ represents the Frobenius scalar product in $\mathbb{R}^{n\times n}$ and $|\cdot|$ denotes its associated norm. The behaviour of the viscoplastic fluids is determined by the magnitud of the stress tensor. The fluid moves as a rigid-solid body if the stress imposed on the fluid is below the yield stress parameter $g \geq 0$. When the stress exceeds g, the fluid deforms and moves as a liquid. Therefore, in the liquid regime, we have that $|\mathcal{E}(\mathbf{u})| \neq 0$, implying continuous deformation. Conversely, if $|\mathcal{E}(\mathbf{u})| = 0$, the material behaves as a solid. Our investigation is focused in the numerical solution of the velocity field $\mathbf{u}$ subject to the incompressibility condition $\operatorname{div}(\mathbf{u}) = 0$.

The first problem considered in this thesis comes from the numerical treatment of the incompressibility constraint. Incompressibility is expressed by the condition $\operatorname{div}(\mathbf{u}) = 0$. Some methods deal with this condition by introducing the pressure as a Lagrange multiplier associated to the velocity divergence-free constraint. Then, the Lagrangian functional is solved by an augmented Lagrangian method [59]. Furthermore, as highlighted in the Motivation, $H(\operatorname{div})$-conforming methods represent a class of finite element methods designed to enforce the exact divergence-free condition for incompressible flows, as detailed in [18, 103]. Particularly, we will cope with incompressibility in a different fashion. In our approach, an exact penalization of the

incompressibility condition is formulated by introducing a nondifferentiable term given by

$$\|\operatorname{div}(\mathbf{u})\|_1 = \int_\Omega |\operatorname{div}(\mathbf{u}(x))|\, dx$$

to penalize the constraint. The L^1-norm is commonly used as an exact penalty for equality constraints [17, Ch. 16]. We seek using an exact penalty to convert a constrained optimization problem into an unconstrained one. Thus, the objective of the first part of this study is to reformulate the original problem as an unconstrained minimization problem with the L^1-norm penalization through the following reformulation:

$$\min_{\mathbf{u}\in\mathbf{H}_0^1(\Omega)} \widetilde{J}(\mathbf{u}) + \sigma\|\operatorname{div}\mathbf{u}\|_{L^1}, \tag{\textbf{EP}}$$

where $\sigma > 0$ is the penalization parameter.

Exact penalty methods rely on the penalty parameter σ, ensuring that the solution of the unconstrained functional coincides with the solution of the constrained problem for all sufficiently large yet finite values of σ [17, 44]. Consequently, a sufficiently large value of σ guarantees the conditions for exactness. In this work we will investigate the exact penalization approach and establish an estimate for σ.

The inclusion of the L^1-norm serves the purpose of promoting sparsity, which enforces the incompressibility condition. However, it entails theoretical and numerical challenges in view of its nonsmoothness. Problem (1.1) inherently includes an additional nondifferentiable term, $\int_\Omega |\mathcal{E}\mathbf{u}|, dx$. In the literature, there are mainly two general strategies to cope with this nonsmooth term: using nonsmooth algorithms or regularization based procedures where the original non-smooth problem is replaced by a smooth approximation of it.

In nonsmooth methods, problem (1.1) has been solved by general augmented Lagrangian methods, the classical *Alternating Direction Method of Multipliers* (ADMM) [52], [5], [59], [49], and an accelerated dual proximal gradient method [112]. On the other hand, the regularization procedure strategy copes with the nondifferentiability involved in the viscoplastic fluid model by regularizing the non-differentiable term, for instance, using a Bercovier-Engelman model [14], the Papanastasiou regularization [97] or a local Huber-C_1 regularization (equivalent to the bi-viscosity model) [37].

Adopting a fully nonsmooth approach to address both nonsmooth terms simultaneously —the exact penalization $\|\operatorname{div}(\mathbf{u})\|_1$ and the term $\int_\Omega |\mathcal{E}\mathbf{u}|\, dx$— result in a convex and nonsmooth problem that can be tackle by several methods proposed in the literature. For instance, the primal-dual proximal splitting (PDPS) of Chambolle and Pock [25] can be applied to the Bingham problem as follows:

$$\min_{\mathbf{u}} \widetilde{J}(\mathbf{u}) + H(\operatorname{div}\mathbf{u}) = \min_{\mathbf{u}} \max_{\boldsymbol{\lambda}} \widetilde{J}(\mathbf{u}) + \langle \operatorname{div}\mathbf{u}, \boldsymbol{\lambda}\rangle - H^*(\boldsymbol{\lambda}),$$

where, for the constraint case, $H = \delta_{\{0\}}$, the $\{0, \infty\}$-valued indicator function, and for the penalty case, $H = \|\cdot\|_{L^1}$.

In this work, we use second-order information and treat these terms separately. Our initial focus is on the incompressibility constraint. To this end, we retain the L^1-norm penalization term and apply the local C_1 Huber regularization to approximate the term $\int_\Omega |\mathcal{E}\mathbf{u}|\, dx$. This regularization approach was previously employed for the Bingham flow in a cylindrical pipe in [37] and for the two-dimensional case in [38]. Accordingly, in the first part of this thesis we will work with a C_1-regularized version of the functional $\widetilde{J}$ while preserving the divergence-free condition.

On the other hand, the second objective of this thesis is to analyze the unregularized functional $\widetilde{J}$. As previously noted, incorporating the nonsmooth penalization $\|\operatorname{div}(\mathbf{u})\|_1$ significantly increases the complexity of the problem. To address this, in the second part of the work we consider the Bingham flow problem in a cylindrical pipe, where the divergence-free condition is automatically satisfied [46, Ch. IV] and the fluid moves just under the effect of the decay of pressure in the pipe. Thus, the velocity field reduces to $\mathbf{u} = (0, 0, u)$. This simplification enables us to focus exclusively on the nonsmooth term $\int_\Omega |\mathcal{E}\mathbf{u}|\, dx$ which is simplified to $\int_\Omega |\nabla u|\, dx$. Thus, problem (1.1) becomes

$$\min_{u \in H_0^1(\Omega)} J(u) := \frac{\mu}{2} \int_\Omega |\nabla u|^2\, dx + g \int_\Omega |\nabla u|\, dx - \int_\Omega cu\, dx, \tag{1.2}$$

where $\Omega \subset \mathbb{R}^2$ corresponds to the cross-section of a pipe and c represents a constant linear decay of preassure.

The unregularized approach to tackle this formulation consists in applying the augmented Lagrangian strategy. To accomplish this, the primal problem (1.2) is reformulated as a linearly constrained minimization problem where the constraint is defined by introducing a new variable $\mathbf{q} = \nabla u$. Thus, the nonsmooth term becomes $\int_\Omega |\mathbf{q}|\, dx$. Moreover, the violation of the constraint is then penalized with an extra quadratic term as follows:

$$\min_{u \in H_0^1(\Omega),\, \mathbf{q} \in [L^2(\Omega)]^2 :\, \mathbf{q} = \nabla u} J(u, \mathbf{q}) := \frac{\mu}{2} \int_\Omega |\nabla u|^2\, dx + g \int_\Omega |\mathbf{q}|\, dx - \int_\Omega cu\, dx + \frac{\rho}{2} \|\nabla u - \mathbf{q}\|^2_{[L^2(\Omega)]^2}. \tag{1.3}$$

Classical algorithms for the numerical solution of this convex optimization problem consists of a set of augmented Lagrangian methods, named as ALG1-ALG4, described in [49]. In our case, in order to solve (1.3) we will use the corresponding saddle-point formulation of the augmented Lagrangian problem and the framework of ALG1, which attempts to find the saddle point by minimizing in the variables $(u, \mathbf{q})$ and

then taking a step along the dual gradient in order to maximize with respect to the Lagrange multiplier [111]. This framework enables the study of the nonsmooth term $\int_\Omega |\mathbf{q}|\, dx$ as a sparsity-promoting regularizer, which enforces zero values for all x in the domain Ω where the material behaves as a rigid solid, i.e., where $|\mathbf{q}(x)| = |\nabla u(x)| = \left|\left(\frac{\partial u(x)}{\partial x_1}, \frac{\partial u(x)}{\partial x_2}\right)\right| = 0$. Consequently, the norm $\int_\Omega |\mathbf{q}(x)|\, dx$, given by

$$\int_\Omega |\mathbf{q}(x)|\, dx = \int_\Omega \left|\left(\frac{\partial u(x)}{\partial x_1}, \frac{\partial u(x)}{\partial x_2}\right)\right| dx = \int_\Omega |(q_1(x), q_2(x))|\, dx = \int_\Omega \sqrt{q_1^2(x) + q_2^2(x)}\, dx,$$

promotes structured sparsity over the vector $\mathbf{q}$ in the solid-like regime. Thus, the pairs $(q_1(x), q_2(x))$ can be analyzed as "groups" and categorized based on whether they exhibit sparsity. Consequently, the discretized version of the term $\int_\Omega |\mathbf{q}(x)|\, dx$ can be associated with the group-sparse norm $\|\cdot\|_{1,2}$, which corresponds to the sum of euclidean norms. We exploit this relation and extend the tools derived for problem (1.3) to a more general formulation in $\mathbb{R}^m$ given by

$$\min_{\mathbf{u}\in\mathbb{R}^m} \psi(\mathbf{u}) := f(\mathbf{u}) + \sigma\|\mathbf{u}\|_{1,2}. \tag{GS}$$

Where f corresponds to a fitting function, typically smooth and not necessarily convex, $\sigma > 0$ corresponds to the penalization parameter, and the $\|\cdot\|_{1,2}$ in finite dimension reads as

$$\|\mathbf{u}\|_{1,2} = \sum_{i=1}^{p} \|\mathbf{u}_i\|_2. \tag{1.4}$$

For $i = 1, \ldots, p$, each $\mathbf{u}_i$ is a subvector of $\mathbf{u}$ called group. Therefore $\mathbf{u}^\top = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p)^\top$. The term $\|\mathbf{u}\|_{1,2}$ promotes sparsity on a group level, i.e., it sets groups of coefficients of $\mathbf{u}$ to zero. In this way, sparsity is promoted groupwise instead of on the individual components of $\mathbf{u}$.

Group sparsity is a highly relevant characteristic of optimization problem solutions, particularly when prior knowledge of the sparsity patterns —often dictated by the applications- is available. In the linear case, the best-known application example is the so-called Group-LASSO problem [9], which consists of minimizing a least-squares fitting term $f(\mathbf{u}) = \frac{1}{2}\|A\mathbf{u} - \mathbf{b}\|^2$ together with the group sparsity term $\|\mathbf{u}\|_{1,2}$. The aim is to select a few groups of variables that serve as predictors in a large classification problem. It was introduced in [9] and [123] to investigate the selection of grouped variables in statistics.

In infinite-dimensional function spaces, particularly in optimal control, directional sparsity [65] is a particular case of group sparsity. In this case, the sparsity term promotes structured sparsity patterns in the optimizing variable or control. In elliptic PDE optimal control problems in two dimension, we may consider $\Omega = \Omega_1 \times \Omega_2$ where

$\Omega_1 = \Omega_2 = (0, 1)$. Thus, the directional sparsity term is given by

$$\|\mathbf{u}\|_{1,2} := \int_{\Omega_1} \left(\int_{\Omega_2} \mathbf{u}^2(x_1, x_2)\, dx_1 \right)^{1/2} dx_2. \tag{1.5}$$

Therefore, the sparsity patterns promoted by this norm are given over the cross sections of the unit square domain.

Moreover, in optimal control problems, (**GS**) is formulated in $L^2(\Omega)$ and given by the tracking type problem:

$$\min_{\mathbf{u} \in L^2(\Omega)} \frac{1}{2}\|S(\mathbf{u}) - \mathbf{y}_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\mathbf{u}\|^2_{L^2(\Omega)} + \sigma\|\mathbf{u}\|_{1,2}, \tag{1.6}$$

where $f(\mathbf{u}) = \frac{1}{2}\|S(\mathbf{u}) - \mathbf{y}_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\mathbf{u}\|^2_{L^2(\Omega)}$ and $S : L^2(\Omega) \to: L^2(\Omega)$ is the control-to-state mapping that assigns to each control $\mathbf{u}$ the unique solution $\mathbf{y}$ of a PDE. It is worth noting that the PDE under consideration may be non-linear, leading to the non-convexity of the function f.

Regarding the numerical solution of problem (**GS**) with f convex, there is a wide range of first-order methods that can be applied to solve it. For instance, the Group-LASSO problem has been addressed using the well-known proximal gradient method, the *Iterative Shrinkage-Thresholding Algorithm* (ISTA), and its accelerated counterpart FISTA [4], the ADMM method [19], and block-coordinate descent methods [100]. In this work, however, we focus on solving (**GS**) with f being non-convex. For the case of non-convex f, a second-order algorithm for sparse L_1-optimization was introduced in [39]. This algorithm employs descent orthantwise directions, introduced by [1], associated with the sparse term $\|\cdot\|_1$. In addition, a projection step is performed to ensure that the iterates remain in the corresponding orthant in order to estimate the active components.

The third objective of this work is to design two numerical algorithms capable of efficiently solve the regularized formulation of the penalized Bingham problem (**EP**) and its simplified, non-regularized version (1.2). In addition, we exploit the ideas developed for the group-sparse term in the non-regularized Bingham problem to design a third efficient algorithm for solving problem (**GS**).

The algorithms proposed in this thesis are built upon a key feature: the integration of generalized second-order information derived from the nonsmooth term. In this thesis, we extend this idea to address the Bingham flow problem and the group-sparse optimization problem. Specifically, we construct a descent direction by solving the minimum norm subgradient problem. To enhance this approach, we aim to accelerate the descent direction by incorporating generalized second-order information. This information is derived from an operator obtained through the generalized differentiation

of the corresponding semismooth function present in the functional.

For the group-sparse optimization problem, an analysis of the optimality condition reveals that the angle between the current iterate and the steepest descent direction must remain non-negative for non-sparse groups in an aproximate solution. Furthermore, with this result we propose an active-set strategy that aims to identify, *a priori*, the sparse groups in an approximate solution. Since the active set, defined as $\mathcal{A}^* := \{i : \mathbf{u}_i^* = \mathbf{0}\}$, at a local minimum $\mathbf{u}^*$ is unknown, our goal is to predict this set at each iteration. Using this prediction, we aim to compute the group-wise steepest descent direction of ψ for problem (**GS**) at every step. In addition, the active-set strategy enables the construction of a reduced second-order matrix that incorporates curvature information exclusively within the inactive groups.

In the following, the organization of this thesis and the main contributions are described in more detail.

Chapter 2 recalls some basic definitions and results concerning convex optimization, non-convex optimization, generalized differentiability, and semismooth functions, which are essential for the development of this work.

Chapter 3 is devoted to the numerical solution of the exact penalization problem (**EP**) for the incompressible Bingham viscoplastic flow. This chapter provides a comprehensive study of the problem, beginning with a review of state-of-the-art. We introduce the steady flow of a Bingham fluid, detailing the constitutive model, and the formulation of the stationary Bingham flow as a variational inequality. In this chapter, we reformulate the Bingham problem addressing both unconstrained and constrained formulations by using the C^1-regularization technique. We review the KKT conditions for the constrained problem. For the unconstrained problem, we explore exact and quadratic penalization methods, emphasizing their role in recovering the fluid's pressure. In addition, we present a second-order method for the exact penalization approach. The algorithm incorporates generalized second-order information, utilizing the notion of semismoothness for superposition operators. We conclude this chapter with extensive numerical experiments, demonstrating the algorithm's performance, comparing it with the semismooth Newton method, and extending its application to 3D geometries.

Chapter 4 explores group-sparsity in the Bingham flow problem and in non-convex optimization. The chapter begins by reviewing the state-of-the-art approaches for the $(l_{1,2})$ norm penalization in both convex and non-convex settings. In addition, Chapter 4 is divided in two sections. In the first one, we analyze the unregularized energy functional associated with the Bingham flow problem in a cylindrical pipe (1.2), where the divergence-free condition is inherently satisfied. Unlike Chapter 3, in this chapter we avoid the C^1-regularization previously applied to the nonsmooth term $g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx$.

Instead, we focus on the augmented Lagrangian formulation of problem (1.2) and the development of an optimization algorithm to solve it. Based on the framework of the well-known ALG1 [49], we propose an algorithm that addresses the inner optimization problem within the ALG1 structure. We study the steepest descent direction and integrate second-order information to enhance the optimization process. We introduce a tailored line-search strategy, followed by an active-set prediction phase to identify and manage rigid zones within the flow. The second-order optimization algorithm is evaluated through numerical experiments and compared against the standard ALG2 method.

In the second part we present a general group-sparse problem formulation and derive the associated optimality conditions. We present several application examples, including nonlinear least-squares problems with group sparsity, elliptic PDE-constrained optimization, and time-dependent PDE-constrained optimization. These examples demonstrate the versatility of group-sparse regularization in diverse contexts. We introduce the Group Sparse Active-set Newton Method (GSNM) and detail its components such as the steepest descent direction, an active-set prediction phase, and the incorporation of second-order information to improve convergence. We analyze the active and inactive index-sets and prove that they are identified at the local minimum. With this result we prove local Q-quadratic convergence of the algorithm. Finally, the chapter addresses the numerical implementation of GSNM and presents computational experiments to validate its performance. Case studies include nonlinear group lasso problems and elliptic optimal control problems, which is compared against the semismooth Newton method.

Contribution of the Thesis

- We analyze the exact penalization problem (**EP**) in the context of nonsmooth optimization. Further, we show the existence of a lower bound for the penalizing parameter that guarantees the equivalence of the divergence penalized optimization problem with the original regularized formulation.

- We show that the fluid's pressure of problem (**EP**) can be recovered from the associated multiplier of the exact penalization by using an analogous approach to the de Rham's theory.

- We propose an algorithm to solve problem (**EP**) which computes a descent direction and is subsequently modified by generalized second–order information. The second–order information is obtained by enriching the Hessian matrix associated with the differentiable part of the cost function $\widetilde{J}$. This enrichment procedure consists of adding a matrix resulting of the generalized differentiation of a Huber

regularization of the L^1 penalizing term of the divergence-free condition.

- We propose a new active-set identification strategy for group-sparse problems. This approach relies on the angle between the current iterate and the steepest descent direction (for each group), enabling rapid identification of the active and inactive sets at the local optimal solution.

- We designed a descent second-order algorithm consisting of three phases to solve group-sparse problems. The first phase involves computing the steepest descent direction with respect to the $\|\cdot\|_{1,2}$ norm. This procedure is performed on a group-wise basis in accordance with the problem's structure. The second phase implements an active-set strategy to predict the active or sparse groups. The third phase incorporates generalized second-order information of the $\|\cdot\|_{1,2}$ norm, in the spirit of [39]. Moreover, the proposed second-order descent algorithm is designed for group-sparse problems that are not necessarily convex.

The majority of the results presented in this thesis are based on previously published or submitted work. Specifically, the key contributions, methodologies, and analyses of Chapters 3 and 4 are derived from [62] and [40] (preprint), where they were first introduced and discussed. While this thesis incorporates these results, it also extends the original work by providing additional context and numerical experiments to enrich the discussion and highlight the broader applicability of the methods.

Published and submitted contribution

Throughout the course of this thesis, we have published or submitted the following contribution:

[62] González-Andrade, López-Ordóñez, and Merino. Nonsmooth exact penalization second-order methods for incompressible bi-viscous fluids. *Computational Optimization and Applications* 80 (2021), 979-1025.

[40] De los Reyes, López-Ordóñez, and Merino. A Second-order Method with Active-set Prediction For Group Sparse Optimization (*preprint*) (2025)

Chapter 2

Preliminaries

This section reviews the definitions, concepts, and fundamental theoretical results that will be used in this thesis. It presents results from both convex and nonconvex optimization, along with the notions of generalized differentiability and semismoothness.

2.0.1 Notation

Along this chapter X, Y and Z are Banach spaces. The notation $\mathcal{H}$ is reserved for Hilbert spaces. The space of bounded linear operators from X to Y is denoted by $\mathcal{L}(X, Y)$. The space $\mathcal{L}(X, \mathbb{R})$ is called the **topological dual space** of X and is denoted by X^*. The duality pairing between X and its dual X^* is given by $\langle \cdot, \cdot \rangle_{X,X^*}$, while any real product defined on X will be denoted by $(\cdot, \cdot)_X$. Moreover, $\| \cdot \|$ represents the norm in the appropriate space, as determined by the context.

2.1 Convex Optimization

This section is based on [11].

Let $\mathcal{H}$ be a Hilbert space and let $f : \mathcal{H} \to [-\infty, +\infty]$. Function f is called **proper** if it never takes the value $-\infty$ and the **effective domain**, $\text{dom } f := \{x \in \mathcal{H} : f(x) < +\infty\}$, is nonempty.

Furthermore, the function f is **convex** if and only if $(\forall x \in \text{dom } f)$ $(\forall y \in \text{dom } f)$ $(\forall \alpha \in]0, 1[)$,

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y),$$

f is **strictly convex** if and only if $(\forall x \in \text{dom } f)$ $(\forall y \in \text{dom } f)$ $(\forall \alpha \in]0, 1[)$,

$$f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha) f(y)$$

whenever $x \neq y$.

Additionally, f is **lower semicontinuous** at $x \in \operatorname{dom} f$ if

$$f(x) \leq \liminf_{n\to\infty} f(x_n)$$

for any sequence $x_n \to x$.

In this work, we also take into account functions that are **Lipschitz continuous**.

Definition 2.1. *Let C be a non-empty and open subset of $\mathcal{H}$, then $f : C \subset \mathcal{H} \to [-\infty, +\infty]$ is Lipschitz continuous on C if there exists a constant $L_f > 0$ so that for all $x, y \in C$,*

$$|f(y) - f(x)| \leq L_f \|y - x\|. \tag{2.1}$$

Moreover, $f : C \subset \mathcal{H} \to [-\infty, +\infty]$ is Lipschitz continuous at $x \in C$ if (2.1) holds for any y in the neighborhood of x. In addition, a function is called locally Lipschitz continuous if for every $x \in C$ there exists a neighborhood U of x such that f restricted to U is Lipschitz continuous.

2.1.1 Differentiability and Subdifferentiability

Let us briefly review the classical notions of differentiability in a Hilbert space.

Let $f : \mathcal{H} \to [-\infty, +\infty]$ be proper, let $x \in \operatorname{dom} f$, and let $y \in \mathcal{H}$. The **directional derivative** of f at x in the direction y is denoted by $f'(x, y)$ and given by

$$f'(x, y) = \lim_{\alpha\to 0^+} \frac{f(x + \alpha y) - f(x)}{\alpha}, \tag{2.2}$$

provided that this limit exists. If this limit exists for all $y \in \mathcal{H}$, then f is called **directionally differentiable** at x in the direction y.

Proposition 2.1. *Let $f : \mathcal{H} \to [-\infty, +\infty]$ be proper and convex, let $x \in \operatorname{dom} f$, and let $y \in \mathcal{H}$. Then $f'(x, y)$ exists in $[-\infty, +\infty]$ and $f'(x, y - x) + f(x) \leq f(y)$.*

Proof. See [11, Prop. 17.2]. □

Let $\mathcal{L}(\mathcal{H}, \mathcal{K})$ be the space of bounded linear operators from $\mathcal{H}$ to the Hilbert space $\mathcal{K}$. Let C be a non-empty and open subset of $\mathcal{H}$ such that $f : C \to \mathcal{K}$. Then, f is **Gâteaux differentiable** at x if there exists an operator $f'(x) \in \mathcal{L}(\mathcal{H}, \mathcal{K})$, called the **Gâteaux derivative** of f at x, such that:

$$(\forall y \in \mathcal{H}),\ \ f'(x)(y) = \lim_{\alpha\to 0} \frac{f(x + \alpha y) - f(x)}{\alpha}.$$

Additionally, if the operator $f'(x) \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ satisfies that

$$\lim_{0 \neq \|y\|_{\mathcal{H}} \to 0} \frac{\|f(u+y) - f(u) - f'(u)(y)\|_{\mathcal{K}}}{\|y\|_{\mathcal{H}}} = 0,$$

we call f **Fréchet differentiable** at u and $f'(u)$ the **Fréchet derivative** of f at u.

The **second Gâteaux derivative** of f at x is the operator $f''(x) \in \mathcal{L}(\mathcal{H}, \mathcal{L}(\mathcal{H}, \mathcal{K}))$ that satisfies:

$$(\forall y \in \mathcal{H}),\ f''(x)(y) = \lim_{\alpha \to 0} \frac{f'(x+\alpha y) - f'(x)}{\alpha}.$$

Similarly, the **second Fréchet derivative** of f at x is the operator $f''(x) \in \mathcal{L}(\mathcal{H}, \mathcal{L}(\mathcal{H}, \mathcal{K}))$ that satisfies:

$$\lim_{0 \neq \|y\|_{\mathcal{H}} \to 0} \frac{\|f'(u+y) - f'(u) - f''(u)(y)\|_{\mathcal{L}(\mathcal{H},\mathcal{K})}}{\|y\|_{\mathcal{H}}} = 0.$$

Let C be a subset of $\mathcal{H}$, let $f : C \to \mathbb{R}$ and suppose that f is Fréchet differentiable, in this case we use the following notation for the Fréchet derivative:

$$f'(x)(y) = \langle f'(x), y \rangle_{\mathcal{H}^*,\mathcal{H}}.$$

Theorem 2.1. ***Riesz-Fréchet representation.*** *Let $h \in \mathcal{L}(\mathcal{H}, \mathbb{R})$. Then there exists a unique vector $u \in \mathcal{H}$ such that $(\forall x \in \mathcal{H})\ h(x) = (x, u)_{\mathcal{H}}$. Moreover, $\|h\| = \|u\|$.*

Proof. See [11, Sec. 2.3]. □

Thanks to the Riesz-Fréchet representation we have that, given $f : C \to \mathbb{R}$ such that f is Fréchet differentiable at $x \in C$, then there exists a unique vector $\nabla f(x) \in \mathcal{H}$, called the **gradient** of f at x, such that

$$(\forall y \in \mathcal{H})\ \langle f'(x), y \rangle_{\mathcal{H}^*,\mathcal{H}} = (\nabla f(x), y)_{\mathcal{H}}. \tag{2.3}$$

Likewise, if f is twice Fréchet differentiable at x, we can identify $f''(x)$ with an operator $\nabla^2 f(x) \in \mathcal{L}(\mathcal{H}, \mathcal{H})$ in the sense that

$$(\forall y \in \mathcal{H})(\forall z \in \mathcal{H})\ \ \langle f''(x)y, z \rangle_{\mathcal{H}^*,\mathcal{H}} = (\nabla^2 f(x)y, z)_{\mathcal{H}}. \tag{2.4}$$

Where $\nabla^2 f(x)$ is called the **Hessian** of f at x.

For $\mathcal{H} = \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient of f at x is defined as

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_n}(x)\right]^\top.$$

Moreover, $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **continuously differentiable** at $x \in \mathbb{R}$, if the partial derivative $\left(\frac{\partial f}{\partial x_i}\right)(x)$ exists and is continuous, for all $i = 1, \cdots, n$.

If f is continuously differentiable at every point of an open set $C \in \mathbb{R}^n$, then f is said to be continuously differentiable on C and denoted by $f \in C^1$.

The next characterization of continuously differentiable functions is a useful result when analyzing convergence properties of optimization algorithms in $\mathcal{H} = \mathbb{R}^n$.

Theorem 2.2. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable in the open convex set $C \subset \mathbb{R}^n$. Let ∇f be Lipschitz continuous at $x \in C$ (with constant $L > 0$). Then for any $x + d \in C$, we have*

$$\|f(x+d) - f(x) - \nabla f(x)d\| \leq \frac{L}{2}\|d\|^2.$$

Proof. By using the mean value theorem in integral form, we have:

$$\begin{aligned} f(x+d) - f(x) - \nabla f(x)d &= \int_0^1 \nabla f(x+td)ddt - \nabla f(x)d \\ &= \int_0^1 (\nabla f(x+td) - \nabla f(x))ddt. \end{aligned}$$

From the locally Lipschitz property of the Jacobian, we get:

$$\begin{aligned} \|f(x+d) - f(x) - \nabla f(x)d\| &\leq \|d\| \int_0^1 \|\nabla f(x+td) - \nabla f(x)\|dt \\ &\leq L\|d\|^2 \int_0^1 tdt \\ &= \tfrac{L}{2}\|d\|^2. \end{aligned}$$

□

Subdifferentiability

For nondifferentiable convex functions, the subdifferential serves as a fundamental tool for their analysis.

Definition 2.2. *Let $f : \mathcal{H} \to [-\infty, +\infty]$ be a convex, proper, and lower semicontinuous function, its **subdifferential** is defined by the set-valued operator given by*

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}} : x \mapsto \{u \in \mathcal{H} : (\forall y \in \mathcal{H})\ \ (y - x, u)_{\mathcal{H}} + f(x) \leq f(y)\}.$$

Here, the notation $\partial f : \mathcal{H} \to 2^{\mathcal{H}}$ means that ∂f maps every point $u \in \mathcal{H}$ to a set $\partial f(u)$ subset of $\mathcal{H}$.

The subdifferential of a convex function can also be characterized by the directional

derivative as follows:

$$\partial f(x) = \{u \in \mathcal{H} : (\forall y \in \mathcal{H}) \ \ (u, y) \leq f'(x, y)\}. \tag{2.5}$$

2.1.2 Existence of Minimizers and Optimality Conditions

We collect the results of sufficient conditions for the existence of minimizers for an extended real-valued convex function and provide a detailed characterization of these minimizers.

Unconstrained case

We are concerned with the convex minimization problem

$$\min_{x \in \mathcal{H}} f(x). \tag{2.6}$$

The following coercivity property is crucial for the existence of minimizers of a convex functional.

A function $f : \mathcal{H} \to [-\infty, +\infty]$ is **coercive** if

$$\lim_{\|x\| \to \infty} f(x) = +\infty.$$

Then, we have the following theorem.

Theorem 2.3. *If $f : \mathcal{H} \to [-\infty, +\infty]$ is proper, convex, coercive, and lower semicontinuous, then problem (2.6) has at least one solution. It has a unique solution if the function f is strictly convex.*

Proof. See [47, Ch. 2 Prop. 1.2]. □

The following principle characterizes global minimizers of a convex proper function.

Theorem 2.4. *(Fermat's Rule) Let $f : \mathcal{H} \to [-\infty, +\infty]$ be convex and proper. Then $\bar{x}$ is a global minimizer of f if, and only if, $0 \in \partial f(\bar{x})$.*

Proof. See [11, Th. 16.2]. □

In addition, if f is Fréchet differentiable at $\bar{x}$, $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$. Then, we have that $\bar{x}$ is a global minimizer of f if, and only if, $\nabla f(\bar{x}) = 0$.

The condition $0 \in \partial f(\bar{x})$, given by Fermat's Rule, is necessary and sufficient (by convexity) for $\bar{x}$ to be a global minimizer.

Constrained case 1

We consider the following constrained convex optimization problem:

$$\begin{aligned} &\min f(x) \\ &\text{s.t } x \in M, \end{aligned} \tag{2.7}$$

where M be a nonempty convex subset of $\mathcal{H}$.

Similarly to the unconstrained case, the existence of minimizers is given by the following result.

Theorem 2.5. *Let $f : \mathcal{H} \to [-\infty, +\infty]$ be proper, convex, and lower semicontinuous. In addition, let us assume that the set M is bounded, or that the function f is coercive over M. Then, the problem (2.7) has at least one solution. It has a unique solution if the function f is strictly convex over M.*

Proof. See [47, Ch. 2 Prop. 1.2]. □

Let us characterize the solutions of problem (2.7).

Theorem 2.6. *Let $f : \mathcal{H} \to [-\infty, +\infty]$ be a proper, lower semicontinuous and convex functional and let M be a nonempty convex subset of dom f. Then, $\bar{x} \in M$ is a solution to the problem (2.7) if, and only if, there exists $u \in \partial f(\bar{x})$ such that $(y - \bar{x}, u)_{\mathcal{H}} \geq 0$ for all $y \in M$. Equivalently, by the characterization given in (2.5), $\bar{x} \in M$ is a solution to (2.7) if, and only if, the directional derivative of f at $\bar{x}$ satisfies that*

$$f'(\bar{x}, y - \bar{x}) \geq 0, \quad \textit{for all } y \in M.$$

Proof. See [11, Prop. 26.5]. □

Additionally, if f is Fréchet differentiable at $\bar{x}$, the necessary and sufficient optimality condition for minimizers is given by the following variational inequality:

$$(\nabla f(\bar{x}), y - \bar{x})_{\mathcal{H}} \geq 0, \quad \text{for all } y \in M.$$

The variational inequality reduces to a variational equation under the following assumptions.

Corollary 2.1. *Assume that $\bar{x} \in \text{int } M$ and $f : M \to [-\infty, +\infty]$ is proper, lower semicontinuous, convex and Gâteaux differentiable at $\bar{x}$. Then, $\bar{x}$ is a solution to the problem (2.7) if, and only if,*

$$(\nabla f(\bar{x}), y)_{\mathcal{H}} = 0, \quad \textit{for all } y \in \mathcal{H}.$$

Proof. See [102, Cor. 5.1.2] □

Theorem 2.7. *(Composite case) Let f_1 and f_2 be lower semicontinuous convex functions of M, with f_2 being differentiable, such that $f = f_1 + f_2$. Then, $\bar{x}$ is a solution of problem (2.7) if, and only if,*

$$-f_2'(\bar{x}) \in \partial f_1(\bar{x}) \Longleftrightarrow f_1(x) - f_1(\bar{x}) + \langle f_2'(\bar{x}), x - \bar{x} \rangle_{\mathcal{H}^*,\mathcal{H}} \geq 0, \text{ for all } x \in M.$$

Proof. See [47, Ch. 2 Prop. 2.2] □

Constrained case 2

For the following kind of constrained optimization problems, necessary and sufficient optimality conditions often require the use of Lagrange multipliers to handle the constraints.

This subsection is based on [74, Ch.1]. Let us consider the constrained problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t} & x \in M \text{ and } g(x) \in K. \end{array} \tag{2.8}$$

Here, we consider that f is a real-valued functional, not necessarily convex, defined on the real Hilbert space $\mathcal{H}$. Moreover, g is a mapping from $\mathcal{H}$ into a real Banach space Z and K is a closed convex cone in Z with vertex at the origin.

Remark 2.1. *Recall that a set K is called **cone** if from $x \in K$ it follows that $\alpha x \in K$ for all $\alpha > 0$.*

The set of all **feasible points** for (2.8) is denoted by $N = \{x \in M : g(x) \in K\} = M \cap g^{-1}(K)$. Further, it is assumed that problem (2.8) admits a solution x^*, i.e., $x^* \in N$ and that f has a local minimum at x^*. Moreover, it is assumed that f is Fréchet differentiable at x^* and g is continuously Fréchet differentiable at x^*.

For a subset A of $\mathcal{H}$, A^+ denotes its **polar cone**:

$$A^+ = \{x^* \in \mathcal{H}^* : \langle x^*, a \rangle_{\mathcal{H}^*,\mathcal{H}} \leq 0 \text{ for all } a \in A\}.$$

Further, for $x \in \mathcal{H}$ the conical hull of $M \backslash \{x\} = M - \{x\}$ is given by

$$M(x) = \{\alpha(c - x) : c \in M, \alpha \geq 0\}.$$

Recall the conical hull is the smallest cone which includes the set $M \backslash \{x\}$.

Definition 2.3. *The Lagrange functional $L : \mathcal{H} \times Z^* \to \mathbb{R}$ associated to problem (2.8) is defined by:*

$$L(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle_{Z^*,Z}.$$

Where $\lambda^* \in Z^*$ is called a Lagrange multiplier for (2.8) at the solution x^* if $\lambda^* \in K^+$, $\langle \lambda^*, g(x^*) \rangle_{Z^*,Z} = 0$, and

$$f'(x^*) + \lambda^* \circ g'(x^*) \in M(x^*)^+,$$

where f' and g' correspond to the Fréchet derivative.

If $M = \mathcal{H}$, then the previous inclusion results in the equation

$$f'(x^*) + \lambda^* \circ g'(x^*) = 0.$$

Lagrange multipliers play a crucial role in formulating both necessary and sufficient optimality conditions for problem (2.8). However, a **regularity condition** might be satisfied in order to ensure the existence of a Lagrange multiplier λ^* satisfying Definition 2.3. Following the classical work of S. Kurcyusz and J. Zowe in [126], the existence of Lagrange multipliers is given if the following condition (see [126, Sec. 1. eq. (1.4)],) is satisfied

$$g'(x^*)M(x^*) - K(g(x^*)) = Z. \tag{2.9}$$

Then, x^* is a called a regular point if (2.9) holds.

With this result we can formulate the necessary optimality conditions known as the **Karush-Kuhn-Tucker** (KKT) conditions for problem (2.8).

Theorem 2.8. *Let x^* be a local solution of (2.8) at which $f : \mathcal{H} \to \mathbb{R}$ and $g : \mathcal{H} \to Z$ are continuously Fréchet differentiable. Assume that the regularity condition or constraint qualification (2.9) is satisfied at x^*.*

Then there exists a Lagrange multiplier $\lambda^* \in Z^*$ such that the Karush-Kuhn-Tucker conditions for (2.8) given by

$$f'(x^*) + \lambda^* \circ g'(x^*) = 0, \tag{2.10a}$$

$$g(x^*) \in K, \tag{2.10b}$$

$$\lambda^* \in K^+ \text{and } \langle \lambda^*, g(x^*) \rangle_{Z^*,Z} = 0 \tag{2.10c}$$

are satisfied.

Proof. See [126, Th. 3.1] □

Example 2.1. *(See [74, Pag. 7]) Suppose that $Z = L^2(\Omega)$ with Ω a domain in $\mathbb{R}^n$ and*

consider the problem

$$\begin{aligned} \min \quad & f(x) \\ s.t \quad & g(x) = 0. \end{aligned} \tag{2.11}$$

Let x^* denote a local solution at which f is Fréchet differentiable and g is continuously Fréchet differentiable. If $g'(x^*) : \mathcal{H} \to Z$ is surjective, the constraint qualification or regularity condition (2.9) is satisfied. Then, there exists $\lambda^* \in Z^*$ such that the KKT condition

$$f'(x^*) + \lambda^* g'(x^*) = 0 \ \text{ in } \ \mathcal{H}^*$$

holds.

2.1.3 Descent Directions

Following the discussion in [68, Pag. 103], determining **descent directions** in a Banach space X is not a standard task. For instance, for a differentiable function f, the derivative $f'(x) \in X^*$ is not appropriate since it is not an element of the space X. However, in a Hilbert space $\mathcal{H}$, by the Riesz representation theorem we can choose $-\nabla f(x) \in \mathcal{H}$ as the steepest descent direction. We follow [11, Sec. 17.4] to define descent directions in the Hilbert space setting .

Recall that the notation $f^{'}(x, d)$ corresponds to the directional derivative of f at x in the direction d.

Theorem 2.9. *Let $f : \mathcal{H} \to]-\infty, +\infty]$ be proper and convex, let $x \in \text{dom } f$. The direction $d \in \mathcal{H}$ is a descent direction for f at x if and only if*

$$f^{'}(x, d) < 0.$$

Proof. [11, Prop. 17.21]. □

Proposition 2.2. *The **steepest descent direction** $\bar{d}$ for a proper and convex function $f : \mathcal{H} \to]-\infty, +\infty]$ is characterized as the solution of the following optimization problem constrained to the unit ball in $\mathcal{H}$:*

$$\min_{\|d\| \leq 1} f'(x, d).$$

Thus, $\bar{d}$ is the unique minimizer of $f'(x, d)$ over all d in the unit ball. Additionally, the steepest descent direction can also be described by the minimum norm subgradient, i.e.,

$$\bar{d} = -\frac{\bar{\eta}}{\|\bar{\eta}\|}, \tag{2.12}$$

where $\bar{\eta}$ is given by:

$$\bar{\eta} = \arg\min_{\eta \in \partial f(x)} \|\eta\|, \tag{2.13}$$

see [11, Prop. 17.22]

2.2 Non-convex optimization

In this section we cover concepts of generalized derivatives for locally Lipschitz continuous functionals. Consider a functional $f : X \to \mathbb{R}$ where X is a Banach space. Recall that f is **locally Lipschitz continuous** near $x \in X$ if there exists a $\delta > 0$ and $L > 0$ such that

$$|f(x_1) - f(x_2)| \leq \|x_1 - x_2\|, \text{ for all } \ x_1, x_2 \in U(x, \delta),$$

where $U(x, \delta)$ is a neighborhood of x

This subsection is based on [32, Ch. 10].

The **Clarke's generalized directional derivative** of $f : X \to \mathbb{R}$ at x in the direction $h \in X$ is given by:

$$f^\circ(x, h) := \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + th) - f(y)}{t}.$$

Unlike the classical definition of the directional derivative, this concept does not require the existence of a limit. Instead, it only considers the behavior of the function f in an arbitrarily small neighborhood around x where the base point y in the difference quotient varies.

Next, the **Clarke subdifferential** for a locally Lipschitz continuous function $f : X \to \mathbb{R}$ at x is defined as

$$\partial_C f(x) = \{\xi \in X^* : \langle \xi, h \rangle_{X^*, X} \leq f^\circ(x, h) \text{ for all } h \in X\}. \tag{2.14}$$

Moreover, we have that:

$$f^\circ(x, h) = \max_{\xi \in \partial_C f(x)} \langle \xi, h \rangle_{X^*, X} \ \text{ for all } h \in X. \tag{2.15}$$

If f is convex and lower semicontinuous, and if $x \in \text{int dom } f$, then $\partial_C = \partial f(x)$.

However, in finite dimensions, an explicit characterization of the Clarke subdifferential can be obtained thanks to Rademacher's Theorem, which is valid only in $\mathbb{R}^n$.

Theorem 2.10. *(Rademacher) Let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$ be Lipschitz continuous. Then f is Fréchet differentiable at almost every $x \in U$.*

Proof. See [45]. □

We denote by $D_f \subset U$ the set of all $x \in U$ at which f admits a Fréchet derivative $f'(x) \in \mathbb{R}^n$. Thus, the characterization of Clarke's subdifferential is given as follows.

Definition 2.4. *Let $U \subset \mathbb{R}^n$ be open and $f : U \to \mathbb{R}$ be Lipschitz continuous near $x \in U$. The set*

$$\partial_B f(x) = \{M \in \mathbb{R}^n : \exists (x_k) \subset D_f : x_k \to x, f'(x_k) \to M\}$$

is called ***Bouligand-subdifferential*** *of f at x. Moreover, Clarke's subdifferential of f at x is the convex hull $\partial_C f(x) = co(\partial_B f(x))$.*

The set $\partial_C f(x) = co(\partial_B f(x))$ is also called **Clarke's generalized Jacobian**.

Theorem 2.11. *Let $U \subset \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be continuously differentiable in a neighborhood of x, then $\partial_C f(x) = \partial_B f(x) = \{f'(x)\}$.*

Proof. See [8, Th. 3.7]. □

Theorem 2.12. *If the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then:*

- *$\partial_C f(x) = \partial f(x)$ and*
- *$f'(x, d) = f^\circ(x, d)$ for all $d \in \mathbb{R}^n$.*

Proof. See [8, Th. 3.8]. □

2.2.1 Optimality Conditions in $\mathbb{R}^n$

Similarly to the convex case, in this section we will review the optimality conditions for the following non-convex problem in $X = \mathbb{R}^n$:

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2.16}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous. The necessary conditions for function f to attain its local minimum are given below.

Theorem 2.13. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $x^* \in \mathbb{R}^n$. If f attains its local minimum at x^*, then*

$$0 \in \partial_C f(x^*) \text{ and } f^\circ(x^*, d) \geq 0, \text{ for all } d \in \mathbb{R}^n.$$

Proof. See [8, Th. 4.1] □

Definition 2.5. *[8, Def. 4.3] A point $x \in \mathbb{R}^n$ satisfying $0 \in \partial_C f(x)$ is called a* ***stationary point*** *of f.*

2.2.2 Descent Directions in Non-convex Optimization

Similarly to the convex case, identifying a direction in which the objective function values decrease is crucial for any descent optimization method. We present a characterization of a descent direction for a locally Lipschitz continuous function.

This section is based on [8, Sec. 4.1.2] and [7, Sec. 1.3].

Theorem 2.14. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $x \in \mathbb{R}^n$. The direction $d \in \mathbb{R}^n$ is a descent direction for f at x if*

$$f^\circ(x, d) < 0.$$

Proof. [8, Sec. 4.1.2]. □

Corollary 2.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $x \in \mathbb{R}^n$. Then either $0 \in \partial_C f(x)$ or there exists a descent direction $d \in \mathbb{R}^n$ for f at x.*

Proof. See [8, Cor. 4.1]. □

Additionally, as discussed in [7, Sec. 1.3], one approach to determining the **steepest descent** direction $\hat{d}$ for f at $x \in \mathbb{R}^n$ is by solving the following optimization problem constrained to the unit ball:

$$\min_{\|d\| \leq 1} f^\circ(x, d), \tag{2.17}$$

which due to (2.15) is equivalent to:

$$\min_{\|d\| \leq 1} \max_{\xi \in \partial_C f(x)} \langle \xi, d \rangle$$

Both the unit ball and $\partial_C f(x)$ are compact sets in $\mathbb{R}^n$, with $\partial_C f(x)$ also being convex. Therefore, by applying von Neumann's minimax theorem, we can interchange the order of optimization:

$$\max_{\xi \in \partial_C f(x)} \min_{\|d\| \leq 1} \langle \xi, d \rangle = \max_{\xi \in \partial_C f(x)} \langle \xi, -\frac{\xi}{\|\xi\|} \rangle = \max_{\xi \in \partial_C f(x)} (-\|\xi\|) = - \min_{\xi \in \partial_C f(x)} \|\xi\|.$$

Thus, this approach is equivalent to finding the generalized subgradient with the minimum norm, i.e., the steepest descent direction is described as

$$\hat{d} = -\frac{\hat{\xi}}{\|\hat{\xi}\|}, \tag{2.18}$$

where $\hat{\xi}$ is given by:

$$\hat{\xi} = \arg\min_{\xi \in \partial_C f(x)} \|\xi\|. \tag{2.19}$$

2.3 Generalized Differentiability and Semismoothness

The concept of **semismoothness** was introduced in [90] for real-valued functions on finite-dimensional spaces, and later extended to vector-valued mappings between finite-dimensional spaces in [98] and [99]. In infinite-dimensional spaces, the notion of semismoothness was extended in [28] and [114]. Moreover, for this type of generalized differentiability the term **slant differentiability** was coined in [28]. Additionally, in [80] and [79], the author introduced a similar notion to slant differentiability and coined the name Newton map.

This section is based on [115].

2.3.1 Semismoothness in Finite-dimensional Spaces

In finite dimensions, for locally Lipschitz continuous functions, semismoothness, as described in [34], involves selecting an appropriate candidate from Clarke's subdifferential, which is conveniently characterized in Definition 2.4 through Rademacher's Theorem. Thus, a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called semismooth at x if f is locally Lipschitz continuous near x and, if for any $s \in \mathbb{R}^n$, the following limit exists:

$$\lim_{\substack{M \in \partial_C f(x+\tau d) \\ d \to s, \tau \to 0^+}} Md.$$

Semismoothness admits equivalent characterizations as follows.

Proposition 2.3. *Let $f : U \to \mathbb{R}$ be defined on the open set $U \subset \mathbb{R}^n$. Then for $x \in U$ the following statements are equivalent:*

- *f is semismooth at x.*
- *f is Lipschitz continuous near x, the directional derivative $f'(x, \cdot)$ exists, and*

$$\sup_{M \in \partial_C f(x+s)} \|Ms - f'(x, s)\| = o(\|s\|) \quad as \ s \to 0.$$

- *f is Lipschitz continuous near x, the directional derivative $f'(x, \cdot)$ exists, and*

$$\sup_{M \in \partial_C f(x+s)} \|f(x+s) - f(x) - Ms\| = o(\|s\|) \quad as \ \ s \to 0.$$

Proof. See [115, Prop. 2.7]. □

Remark 2.2. *Directional differentiability and Bouligand differentiability are equivalent for locally Lipschitz continuos mappings between finite-dimensional spaces [104].*

Example 2.2. *(*max *function [67, Lemma 3.1]) The mapping* $y \to \max(0, y)$ *from* $\mathbb{R}^n$ *to* $\mathbb{R}^n$ *is semismooth or slantly differentiable on* $\mathbb{R}^n$*, and the following matrix-valued function* $G_m : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ *is the slant or generalized derivative defined by*

$$G_m(y) = \ diag\ (g_1(y_1), \cdots, g_n(y_n)),$$

where $g_i : \mathbb{R} \to \mathbb{R}$ *is given by*

$$g_i(z) = \begin{cases} 0 & \text{if } z < 0, \\ 1 & \text{if } z > 0, \\ \delta_i & \text{if } z = 0, \end{cases} \tag{2.20}$$

with $0 \leq \delta_i \leq 1$.

Additionally, the composition of semismooth functions is semismooth [115, Prop. 2.9].

Proposition 2.4. *Let* $U \subset \mathbb{R}^n$ *and* $V \subset \mathbb{R}^l$ *be open sets. Let* $g : U \to V$ *be semismooth at* $x \in U$ *and* $h : V \to \mathbb{R}^m$ *be semismooth at* $g(x)$ *with* $g(U) \subset V$*. Then the composite map* $f = h \circ g : W \to \mathbb{R}^m$ *is semismooth at* x*.*

Proof. See [115, Prop. 2.9]. □

2.3.2 Semismoothness in Infinite-dimensional Spaces

In this section, we review the concept of semismoothness in Banach spaces for superposition operators.

Superposition operators or **Nemytskii operators** are a class of operators in L^p spaces that result from the composition of L^p - nonlinear functions.

Following the work and notation introduced in [115], we consider Nemytskii (or superposition) operators $\Phi : Y \to L^r(\Omega)$, defined by

$$\Phi(u)(x) = \phi(F(u)(x)) \tag{2.21}$$

for almost all x on Ω. Here, Y is a real Banach space. The mappings $\phi : \mathbb{R}^m \to \mathbb{R}$ and $F : Y \to \prod_{i=1}^m L^{r_i}(\Omega)$, with $1 \leq r \leq r_i < \infty$, satisfy the following conditions [115][Assump. 3.32]:

Assumption 2.1. *There are $1 \leq r \leq r_i < q_i \leq \infty$, for $1 \leq i \leq m$, such that:*

a) The mapping $F : Y \to \prod_{i=1}^m L^{r_i}(\Omega)$ is continuously Fréchet differentiable.

b) The mapping $Y \ni u \mapsto F(u) \in \prod_{i=1}^m L^{q_i}(\Omega)$ is locally Lipschitz continuous, i.e., for all $u \in Y$ there exists an open neighborhood $V(u)$ and a constant $L_F(V)$ such that:

$$\sum_i \|F_i(u_1) - F_i(u_2)\|_{L^{q_i}} \leq L_F \|u_1 - u_2\| \ \ \forall u_1, u_2 \in V.$$

c) The function $\phi : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuos, i.e.,

$$|\phi(x_1) - \phi(x_2)| \leq L_\phi \|x_1 - x_2\|, \ \forall x_1, x_2 \in \mathbb{R}^m.$$

d) ϕ is semismooth.

In addition, we use the following definition of semismoothness [115][Def. 3.48].

Definition 2.6. *The operator Φ is called semismooth at $y \in Y$ if it satisfies*

$$\sup_{G \in \partial^\circ \Phi(u+h)} \|\Phi(u+h) - \Phi(u) - Gh\|_{L^r} = o(\|h\|_Y), \quad \text{as } h \to 0 \text{ in } Y, \tag{2.22}$$

where $\partial^\circ \Phi$ corresponds to the generalized differential:

$$\partial^\circ \Phi(u) = \left\{ \begin{array}{l} G \in \mathcal{L}(Y, L^r) \text{ such that } G : v \mapsto \sum_i M_i(u)(F_i'(u)v), \\ \text{where } M(u) \text{ is a measurable selection of Clarke's generalized Jacobian } \partial_C \phi(F(u)) \end{array} \right\} \tag{2.23}$$

We conclude this section with the following result about semismoothness of superposition opeators.

Theorem 2.15. *Under Assumption 2.1, the operator (2.21) is semismooth on Y.*

Proof. See [115, Th. 3.49] □

Chapter 3

Part I: Exact Penalty Approach for the Incompressibility Condition in the Bingham Flow Problem

This chapter focuses on the analysis and design of a second-order optimization algorithm for the numerical solution of the stationary flow of a Bingham fluid. The algorithm is designed to address a penalized nonsmooth energy functional associated with the following Bingham flow problem:

$$\begin{cases} \min\limits_{\mathbf{u}\in\mathbf{H}_0^1(\Omega)} \widetilde{J}(\mathbf{u}) := \mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{u}\, dx + g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx - \int_\Omega \mathbf{f_b} \cdot \mathbf{u}\, dx, \\ \text{s.t.} \quad \operatorname{div} \mathbf{u} = 0, \end{cases}$$

where $\mathbf{u}$ represents the velocity field and $g > 0$ corresponds to the yield stress parameter.

To address the incompressibility condition, $\operatorname{div} \mathbf{u} = 0$, we employ an exact penalization in terms of the L^1-norm. This penalization transforms the constraint problem into a nonsmooth unconstrained optimization formulation:

$$\min_{\mathbf{u}\in\mathbf{H}_0^1(\Omega)} \widetilde{J}(\mathbf{u}) + \sigma \| \operatorname{div} \mathbf{u}\|_1,$$

where $\sigma > 0$ is the penalization parameter. For this formulation, we propose a second-order descent algorithm. Our method solves the resulting nonsmooth problem by considering the steepest descent direction and extra generalized second–order information associated to the nonsmooth term. This method has the advantage that the divergence-free property is enforced by the designed descent direction without the need of build-in divergence-free approximation schemes.

In Section 3.1, we review the state-of-the-art approaches for the Bingham flow prob-

lem. Section 3.2 focuses on the constitutive model for the steady flow of a Bingham fluid and the formulation of this problem as a variational inequality. Since the variational inequality serves as the optimality condition for a convex optimization problem, we present the constrained convex problem and its regularized formulation using the C^1-Huber regularization. In Section 3.3.2, we examine the KKT conditions for the constrained problem.

Section 3.4 analyzes the exact penalization of the divergence-free constraint and provides a sharp estimation of the penalty parameter required to achieve exact penalization. Quadratic penalization is discussed in Section 3.5. Additionally, Section 3.6 addresses the recovery of the fluid's pressure for both penalty methods—quadratic and exact penalizations. In this section, we argue the existence of a function that serves as the pressure of the fluid.

The subsequent sections focus on the analysis of a second-order method for the exact penalization approach. This algorithm incorporates generalized second-order information, utilizing the notion of semismoothness for superposition operators. The chapter concludes with numerical experiments comparing it to the Semismooth Newton method and extending its application to 3D geometries.

The main results in this Chapter were first obtained in a joint work with Sergio González and Pedro Merino in [62].

3.1 State-of-the-Art: Optimization Methods for Solving Bingham Fluids

The numerical simulation of incompressible Bingham fluids has advanced significantly with the application of various optimization techniques. This progress is rooted in the fact that the flow of a Bingham fluid is characterized by a variational inequality of the second kind, which serves as the optimality condition for a nondifferentiable convex optimization problem.

The optimization methods address the inherent challenges posed by the yield stress behavior of Bingham fluids, which exhibit both solid and liquid characteristics depending on the applied stress.

This review summarizes the state-of-the-art techniques used in the simulation of Bingham fluids, focusing on optimization methods.

One of the classical approaches for minimizing the Bingham energy functional is the Uzawa algorithm [56, 110], which is applied to the equivalent saddle-point problem. A closely related technique is the Augmented Lagrangian method [49, 52], which introduces an additional multiplier interpreted as the extra stress tensor. Both meth-

ods are particularly effective in accurately capturing the transition between yielded and unyielded regions. However, for large-scale problems, the need for detailed mesh refinement can lead to increased computational demands, which makes standard numerical discretization challenging [119, 124]. Numerous studies in the literature have implemented algorithms based on the Augmented Lagrangian framework, including [42, 49, 50, 71, 72, 91, 92, 101, 117].

In [16], the authors propose an alternative optimization method to traditional penalization and augmented Lagrangian techniques, which consists in the use of second-order cone programming (SOCP). This approach formulates the minimum principle for Bingham fluid flows as an SOCP problem, which can be efficiently solved using primal-dual interior point solvers. This method avoids the need for regularization and mixed stress-velocity approaches.

In [38], the authors proposed a semismooth Newton Method for the numerical simulation of two-dimensional stationary Bingham fluid flow. This method involves the Tikhonov regularization and the use of Fenchel's duality to obtain optimality systems. The proposed semismooth Newton algorithm is formulated in finite dimensions and incorporates additional regularization, ensuring local superlinear convergence and delivering efficient numerical performance.

The integration of the augmented Lagrangian method with physics-informed neural networks (PINNs) was proposed in [125] for the Bingham fluid flow simulation. This method uses the flexibility of PINNs, allowing for the learning of parameter-dependent numerical solutions through the training process. The augmented Lagrangian method used in conjunction with PINNs provides a feasible loss function for deep learning, enhancing the simulation's accuracy and efficiency. However, this method can lead to large-scale,ill-conditioned linear systems that require preconditioning. This adds computational overhead and complexity to the simulation process.

Regularization techniques continue to play a crucial role in stabilizing numerical simulations, ensuring accurate and reliable results. For instance, regularization such as the Papanastasiou model [97], are commonly used to address the computational difficulties associated with the classical Bingham model as in [107]. These techniques smooth out the yield stress behavior, making the numerical simulation more stable.

Additionally, enforcing the incompressibility or divergence-free condition in numerical simulations presents a significant challenge. Below is a summary of the approaches employed to address this issue.

Several advancements have been made in the Smoothed Particle Hydrodynamics (SPH) framework to handle the incompressibility condition. For instance, in [13] the Divergence-Free SPH method enforces incompressibility by maintaining a constant density and a divergence-free velocity field.

If the finite–element–method is used for the numerical approximation of viscoplastic fluids, several important stability issues arise that need to be addressed. Particularly, the well–known Ladyzhenskaya–Babuška–Brezzi condition [55, Ch.2 Sec.1.4] must hold in order to guarantee a stable approximation. This conditions is satisfied, for instance, by Taylor–Hood finite elements [55, Ch.2 Sec.4.2]. In addition, the divergence–free constraint is another important condition which can be incorporated in the finite element approximation. In [26], a numerical model for the simulation of incompressible two-phase flows is presented. Here, a finite element method combined with a penalization method has been used to handle the incompressibility constraint in two-phase flow simulations.

3.2 Steady Flow of a Bingham Fluid

3.2.1 Notation

The following notation will be used thorough the rest of the chapter. The Euclidean norm in $\mathbb{R}^n$ is denoted by $|\cdot|$. The duality pairing between a Banach space Y and its dual Y^* is given by $\langle\cdot,\cdot\rangle_{Y,Y^*}$, while any real inner product defined on Y will be noted by $(\cdot,\cdot)_Y$.

Recall that a domain Ω is said to have a Lipschitz boundary if its boundary can be locally represented as the graph of a Lipschitz continuous function. Accordingly, we assume that the domain Ω is an open and bounded subset of $\mathbb{R}^n$, for $n = 2, 3$, with Lipschitz boundary.

The Frobenius scalar product in $\mathbb{R}^{n\times n}$ and its associated norm are defined by

$$A : B = \operatorname{tr}(AB^\top) \quad \text{and} \quad |A| = \sqrt{(A : A)}, \quad \text{for } A,\, B \in \mathbb{R}^{n\times n},$$

respectively. Recall that, for simplicity of notation, both the Euclidean norm and the Frobenius norm are denoted by $|\cdot|$ Any potential ambiguity is avoided due to the context provided by the argument.

We use the following spaces: $L^2(\Omega)$ is the collection of square-integrable functions defined on Ω,

$$\begin{aligned}
H^1(\Omega) &= \left\{ u \in L^2(\Omega) : \frac{\partial u}{\partial x_i} \in L^2(\Omega) \ \text{ for } i = 1, \cdots, n \right\},\\
H^1_0(\Omega) &= \left\{ u \in H^1(\Omega) : u|_\Gamma = 0 \right\},\\
L^2_0(\Omega) &= \left\{ q \in L^2(\Omega) : \int_\Omega q\, dx = 0 \right\}.
\end{aligned}$$

We use the bold notation for the vector spaces, such as $\mathbf{H}_0^1(\Omega) = (H_0^1(\Omega))^n$. Further, we introduce the space of symmetric matrices of L^p-functions as

$$\mathbb{L}^p(\Omega) := \left\{ \boldsymbol{\tau} = (\tau_{ij})_{i,j=1}^n \,:\, \tau_{ij} = \tau_{ji} \in L^p(\Omega) \right\}.$$

Taking into account that $\mathbf{u}$ is the velocity field of a fluid, the divergence–free space is given by

$$V = \{\mathbf{u} \in \mathbf{H}_0^1(\Omega) \,:\, \operatorname{div} \mathbf{u} = 0\}.$$

Finally, we use the notation p_r for the pressure and $\mathcal{E} = \frac{1}{2}(\nabla + \nabla^\top)$ for the symmetric gradient operator - strain rate tensor. Considering that $\mathcal{E} : \mathbf{H}_0^1(\Omega) \to \mathbb{L}^2(\Omega)$, we obtain that $\mathcal{E}\mathbf{u} = (\mathcal{E}_{ij}(\mathbf{u}))$, with $\mathcal{E}_{ij}(\mathbf{u}) := \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}) \in L^2(\Omega)$ for $1 \leq i, j \leq n$. Moreover, by [31, pp. 404], we have that $\displaystyle\int_\Omega \mathcal{E}\mathbf{u}(x) : \mathcal{E}\mathbf{u}(x)\, dx = \|\mathcal{E}\mathbf{u}\|^2_{\mathbb{L}^2}$.

3.2.2 Constitutive Model

Viscoplastic fluids are materials distinguished by a yield stress, meaning they flow like liquids when the applied stress surpasses the yield stress; otherwise, they act as rigid solids [70]. In particular, we focus on Bingham fluids, which are a specific class of yield stress fluids.

Precise constitutive equations exist to characterize viscoplastic fluids, enabling a clear distinction between yielded and unyielded regions. The differentiation between zones is determined by the relationship between stress and the rate of strain tensor, which describes the relative deformation of a material and is defined in terms of the symmetric components of the velocity gradient. We will review the constitutive relation of the Bingham material involving the stress, denoted by τ, and the rate of strain tensor given by $\mathcal{E}\mathbf{u} = \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^\top)$. When the rate of strain is zero, the magnitud of the stress is below the yield stress parameter $g \geq 0$ and the fluid moves as a rigid-solid body. When the stress exceeds g, the fluid deforms and moves as a liquid. This dual behavior is modeled by

$$\begin{cases} |\tau| \leq g, & if\ \mathcal{E}\mathbf{u} = 0, \\ \tau = 2\mu\mathcal{E}\mathbf{u} + g\frac{\mathcal{E}\mathbf{u}}{|\mathcal{E}\mathbf{u}|}, & if\ \mathcal{E}\mathbf{u} \neq 0, \end{cases} \tag{3.1}$$

where $\mu > 0$ is the viscosity of the fluid.

Additionally, the strong formulation of the Bingham flow problem is based on the principles of momentum and mass conservation described by the Navier-Stokes equations. These equations, combined with the rheological model defined by (3.1) and appropriate boundary conditions form the complete mathematical model of the Bingham flow problem. We analyze the stationary flow model of a Bingham fluid, where

the functions involved in the model are independent of time. Let Ω be an open set of $\mathbb{R}^n$, $n \in \{2, 3\}$, with regular boundary Γ. The flow equations for a Bingham fluid in Ω are given by the following system of equations (see [60] and the references therein):

$$\begin{cases} \text{Div } \tau + \nabla p_r + \mathbf{f_b} = \mathbf{0}, & in\ \Omega, \\ \text{div } \mathbf{u} = 0, & in\ \Omega, \\ |\tau| \leq g, & if\ \mathcal{E}\mathbf{u} = 0, \\ \tau = 2\mu\mathcal{E}\mathbf{u} + g\frac{\mathcal{E}\mathbf{u}}{|\mathcal{E}\mathbf{u}|}, & if\ \mathcal{E}\mathbf{u} \neq 0, \\ + \text{ Boundary conditions} & on\ \Gamma. \end{cases} \tag{3.2}$$

Here, Div denotes the row-wise divergence operator, $\mathbf{f_b}$ represents the body force acting on the fluid, and p_r corresponds to the fluid's pressure.

3.2.3 Stationary Bingham Flow as a Variational Inequality

In the foundational work [46], the weak formulation of system (3.2) was established. Consequently, our focus is on the steady Bingham flow (3.2) expressed in its variational form. In the stationary case, we assume the fluid flows under the influence of a forcing term $\mathbf{f_b}$ and is confined within the domain Ω. Consequently, we impose homogeneous Dirichlet boundary conditions.

Let us define the function space for the variational formulation as follows:

$$V = \{\mathbf{u} \in \mathbf{H}_0^1(\Omega) \,:\, \text{div}\, \mathbf{u} = 0\}.$$

We refer to $\widetilde{\mathbf{u}} \in V$ as the weak solution of the stationary viscoplastic flow problem (3.2) if the following variational inequality of the second kind holds (see [46], [110], [57], and the references therein):

$$a(\mathcal{E}\widetilde{\mathbf{u}}, \mathcal{E}\mathbf{v} - \mathcal{E}\widetilde{\mathbf{u}}) + j(\mathcal{E}\mathbf{v}) - j(\mathcal{E}\widetilde{\mathbf{u}}) \geq \langle \mathbf{f_b}, \mathbf{v} - \widetilde{\mathbf{u}} \rangle_{\mathbf{L}^2(\Omega)} \quad \text{for all } \mathbf{v} \in V, \tag{3.3}$$

where the bilinear form $a(\cdot, \cdot)$ is given by:

$$a(\mathcal{E}\mathbf{u}, \mathcal{E}\mathbf{v}) := 2\mu \int_\Omega \mathcal{E}\mathbf{u}(x) : \mathcal{E}\mathbf{v}(x)\, dx,$$

and

$$j(\mathcal{E}\mathbf{u}) := g\sqrt{2} \int_\Omega |\mathcal{E}\mathbf{u}(x)|\, dx.$$

Moreover, the variational inequality (3.3) represents an optimality condition for a convex optimization problem. From Theorem 2.7, we obtain the following characterization of the weak solution: the velocity field $\widetilde{\mathbf{u}} \in V$ is a weak solution of the stationary

Bingham problem, if, and only if,

$$\widetilde{\mathbf{u}} = \arg\min_{\mathbf{u}\in V} \widetilde{J}(\mathbf{u}) := \mu \int_\Omega \mathcal{E}\mathbf{u}(x) : \mathcal{E}\mathbf{u}(x)\, dx + g \int_\Omega |\mathcal{E}\mathbf{u}(x)|\, dx - \int_\Omega \mathbf{f_b}(x) \cdot \mathbf{u}(x)\, dx. \quad (3.4)$$

For clarity and brevity, the dependence on x in the velocity field $\mathbf{u}$ will be omitted in the following sections, except where explicitly required to prevent ambiguity.

3.3 Bingham Problem as a Convex Optimization Problem

3.3.1 Unconstrained Problem and Regularization

From the last section we deduced that the solution of the following nondifferentiable convex problem corresponds to the velocity field of the steady-state Bingham flow.

$$\min_{\mathbf{u}\in V} \widetilde{J}(\mathbf{u}) := \mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{u}\, dx + g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx - \int_\Omega \mathbf{f_b} \cdot \mathbf{u}\, dx. \quad (3.5)$$

The nondifferentiability in problem (3.5) arises from the norm in the second term, $g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx$. We address this nonsmoothness by employing a local regularization technique.

Huber regularization

The Huber regularization function [69] is a widely used tool in optimization and statistical modeling, particularly for robust regression and regularization techniques. The Huber function is controlled by a threshold or approximation parameter, $\beta > 0$, that determines the point of transition between a quadratic form for small values of the argument and a linear form for large values. The Huber regularization for the absolute value $|\cdot|$ is defined by a piecewise function $\psi : \mathbb{R} \to \mathbb{R}$ given by:

$$\psi(u) = \begin{cases} |u| - \frac{1}{2\beta}, & if\ |u| \geq \frac{1}{\beta}, \\ \frac{\beta}{2}|u|^2, & if\ |u| < \frac{1}{\beta}. \end{cases}$$

In Figure 3.1, we illustrate that as β increases, the approximation of the absolute value becomes more accurate.

The Huber regularization for the Bingham flow problem, introduced in [37], represents a local C^1 regularization of the Frobenius norm, providing a smooth approxima-

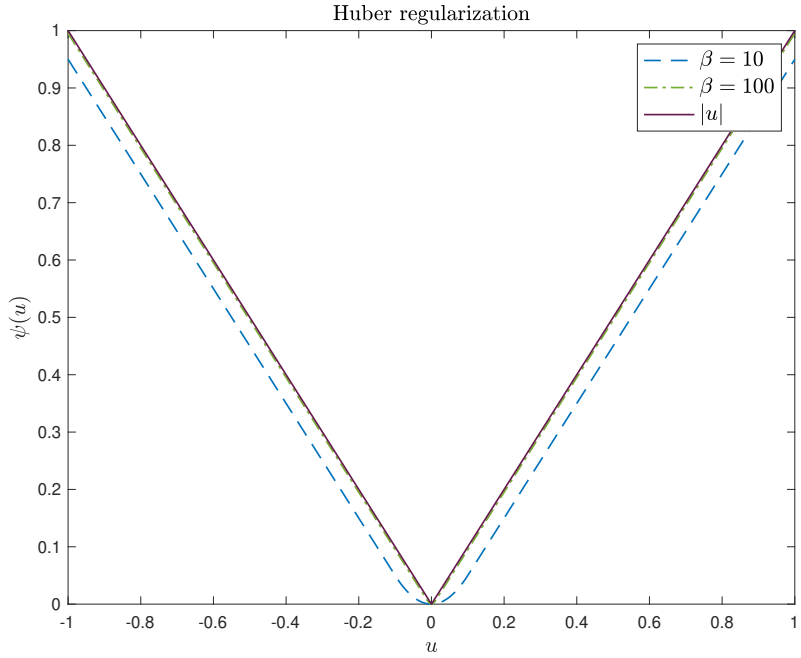


Figure 3.1: Absolute value's Huber regularization.

tion that facilitates numerical computations and analysis. It is denoted by $\Psi : \mathbb{R}^{d\times d} \to \mathbb{R}$ and defined by:

$$\Psi(A) = \begin{cases} g|A| - \frac{g^2}{2\beta}, & if\ |A| \geq \frac{g}{\beta}, \\ \frac{\beta}{2}|A|^2, & if\ |A| < \frac{g}{\beta}, \end{cases} \tag{3.6}$$

where $\beta > 0$ is approximation parameter. Clearly, we have that $\beta \to \infty$ implies that $\Psi(A) \to |A|$.

Thus, the regularized Bingham problems is given by

$$\min_{\mathbf{u}\in V} J(\mathbf{u}) := \mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{u}\, dx + \int_\Omega \Psi(\mathcal{E}\mathbf{u})\, dx - \int_\Omega \mathbf{f_b} \cdot \mathbf{u}\, dx. \tag{3.7}$$

Bi-viscosity Model

The regularized problem (3.7) can be equivalently derived from a smooth approximation of the Bingham constitutive laws (3.1). Specifically, this smooth approximation is obtained using the bi-viscosity model, as introduced in [108], [95], and [51]. This bi–viscous fluid rheology is described by introducing, for the solid regime, a second viscosity μ_1 which is large but finite. In the liquid-like regime, the original Bingham model holds with the viscosity μ bounded. A dimensionless regularization parameter ϵ, which is a function of the ratio between μ and μ_1, is also introduced in the bi-viscous model. Therefore, the stress in the bi–viscous rheology can be written as follows (see

[111, Sec. 4.30]):

$$\tau = 2\mu\mathcal{E}\mathbf{u} + g\frac{\mathcal{E}\mathbf{u}}{\max(\frac{\epsilon}{\mu}g, |\mathcal{E}\mathbf{u}|)}. \tag{3.8}$$

Thus, bi-viscosity approach provides an equivalent formulation as the one obtained by the Huber regularization given in (3.7). Then, parameter β in (3.6) can be approximated in terms of $1/\epsilon$, where ϵ is the bi-viscosity model parameter.

Uniqueness of the solution

The objective funcional J in (3.7) is proper, strictly convex, continuous and coercive over $V \subset \mathbf{H}_0^1(\Omega)$ (see [56, 60, 86]). Therefore, from the review of convex analysis in Chapter (2), Theorem (2.5) implies the existence of a unique solution $\tilde{\mathbf{u}} \in V \subset \mathbf{H}_0^1(\Omega)$ of (3.7).

The results presented in the following sections were first introduced and proved in [62].

3.3.2 Constrained Problem and KKT Conditions

The unconstrained optimization problem (3.7) is posed in the divergence–free space V. Thus, we can reformulate this problem as the following constrained optimization problem in $\mathbf{H}_0^1(\Omega)$:

$$\begin{cases} \min\limits_{\mathbf{u}\in\mathbf{H}_0^1(\Omega)} J(\mathbf{u}) := \mu\int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{u}\, dx + \int_\Omega \Psi(\mathcal{E}\mathbf{u})\, dx - \int_\Omega \mathbf{f_b}\cdot\mathbf{u}\, dx \\ \text{subject to:} \quad \operatorname{div}\mathbf{u} = 0, \end{cases} \tag{CP}$$

where $\operatorname{div} : \mathbf{H}_0^1(\Omega) \to L_0^2(\Omega)$. (CP) is a convex differentiable optimization problem with differentiable equality constraints that has a unique solution.

A regularity condition is needed in order to guarantee the existence of a Lagrange multiplier that allows us to derive a Karush-Kuhn-Tucker (KKT) system for the constrained problem (CP). Thus, following the review presented in Section 2.1.2, we need to prove that the regularity condition (2.9) holds (see Theorem 2.8). In our context, since the constraint is given by the linear operator div, and its derivative is the operator itself, $\operatorname{div} : \mathbf{H}_0^1(\Omega) \to L_0^2(\Omega)$, condition (2.9) takes the following form

$$C = Z = L_0^2(\Omega), \tag{3.9}$$

where $C := \{\alpha \operatorname{div}\mathbf{v} : \alpha \geq 0, \mathbf{v} \in \mathbf{H}_0^1(\Omega)\}$ is the cone generated by the image of the divergence operator. Moreover, from Example 2.11, an immediate observation is that the condition (3.9) is satisfied since the continuous linear operator - $\operatorname{div} : \mathbf{H}_0^1(\Omega) \to$

$L^2_0(\Omega)$ - is surjective (see [31, Th. 6.14-1]). Therefore, by Theorem 2.8, we infer the existence of a Lagrange multiplier $\lambda \in L^2_0(\Omega)$ associated to the constraint $\operatorname{div} \mathbf{u} = 0$, such that the following system at the solution $\widetilde{\mathbf{u}}$, is satisfied:

$$\operatorname{div} \widetilde{\mathbf{u}} = 0, \tag{3.10a}$$

$$\begin{aligned} \langle J'(\widetilde{\mathbf{u}}) + \operatorname{grad} \lambda, \mathbf{u} - \widetilde{\mathbf{u}} \rangle_{\mathbf{H}^{-1}, \mathbf{H}^1_0} &= \langle J'(\widetilde{\mathbf{u}}), \mathbf{u} - \widetilde{\mathbf{u}} \rangle_{\mathbf{H}^{-1}, \mathbf{H}^1_0} \\ &- (\lambda, \operatorname{div}(\mathbf{u} - \widetilde{\mathbf{u}}))_{L^2} = 0, \ \ \forall \, \mathbf{u} \in \mathbf{H}^1_0(\Omega). \end{aligned} \tag{3.10b}$$

Here, $-\operatorname{div} : \mathbf{H}^1_0(\Omega) \to L^2_0(\Omega)$ is the dual operator of $\operatorname{grad} : L^2_0(\Omega) \to \mathbf{H}^{-1}(\Omega)$ (see [31, Th. 6.14-1]) and $J'(\mathbf{u}) : \mathbf{H}^1_0(\Omega) \to \mathbb{R}$ is the Fréchet derivative of J at $\mathbf{u}$, given by

$$\langle J'(\mathbf{u}), \mathbf{v} \rangle_{\mathbf{H}^{-1}, \mathbf{H}^1_0} = \ 2\mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v} + g \int_\Omega \beta \frac{\mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v}}{\max(g, \beta|\mathcal{E}\mathbf{u}|)} \, dx - \int_\Omega \mathbf{f_b} \cdot \mathbf{v} \, dx. \tag{3.11}$$

Recall that $L^2_0(\Omega)$ is identified with its dual space.

Notice that the last derivative has a nondifferentiable term involving the max function.

The following technical result regarding this nondifferentiable term will be useful in the forthcoming analysis.

Lemma 3.1. *Let $g > 0$ be given. For fixed $\beta > 0$, let $\mathbf{u} \in \mathbf{H}^1_0(\Omega)$ and let $E^{\mathbf{u}}_\beta := \{x \in \Omega \, : \, \beta|\mathcal{E}\mathbf{u}| < g\}$ and $I^{\mathbf{u}}_\beta := \{x \in \Omega \, : \, \beta|\mathcal{E}\mathbf{u}| \geq g\}$. The superposition operator $\theta_\beta(\mathbf{u}) := \max(g, \beta|\mathcal{E}\mathbf{u}|)$ satisfies the following inequalities:*

$$\theta_\beta(\mathbf{u})(x) - \theta_\beta(\mathbf{v})(x) \begin{cases} = 0, & \text{if } \ x \in E^{\mathbf{u}}_\beta \cap E^{\mathbf{v}}_\beta, \\ \leq 0, & \text{if } \ x \in E^{\mathbf{u}}_\beta \cap I^{\mathbf{v}}_\beta, \\ \leq \beta|\mathcal{E}\mathbf{u}(x) - \mathcal{E}\mathbf{v}(x)|, & \text{if } \ x \in E^{\mathbf{v}}_\beta \cap I^{\mathbf{u}}_\beta, \\ \leq \beta|\mathcal{E}\mathbf{u}(x) - \mathcal{E}\mathbf{v}(x)|, & \text{if } \ x \in I^{\mathbf{u}}_\beta \cap I^{\mathbf{v}}_\beta, \end{cases} \tag{3.12}$$

for almost all x in Ω and for all $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{H}^1_0(\Omega)$.

Proof. Let $\mathbf{u}, \mathbf{v} \in \mathbf{H}^1_0(\Omega)$. We analyse pointwise bounds of $\theta_\beta(\mathbf{u}) - \theta_\beta(\mathbf{v})$ on the four disjoint sets: $E^{\mathbf{u}}_\beta \cap E^{\mathbf{v}}_\beta$, $E^{\mathbf{u}}_\beta \cap I^{\mathbf{v}}_\beta$, $E^{\mathbf{v}}_\beta \cap I^{\mathbf{u}}_\beta$ and $I^{\mathbf{u}}_\beta \cap I^{\mathbf{v}}_\beta$.

Consider $E^{\mathbf{u}}_\beta \cap E^{\mathbf{v}}_\beta$. In this set, (3.12) is directly satisfied since $\theta_\beta(\mathbf{u}) = \theta_\beta(\mathbf{v}) = g$. Hence $\theta_\beta(\mathbf{u}) - \theta_\beta(\mathbf{v}) = 0$. On $E^{\mathbf{u}}_\beta \cap I^{\mathbf{v}}_\beta$, we have that $\theta_\beta(\mathbf{u}) = g$ and $\theta_\beta(\mathbf{v}) = \beta|\mathcal{E}\mathbf{v}|$ with $\beta|\mathcal{E}\mathbf{v}| \geq g$. Therefore, we have the relation:

$$\theta_\beta(\mathbf{u}) - \theta_\beta(\mathbf{v}) = g - \beta|\mathcal{E}\mathbf{v}| \leq g - g = 0 \leq \beta|\mathcal{E}\mathbf{u} - \mathcal{E}\mathbf{v}|. \tag{3.13}$$

Next, in $E^{\mathbf{v}}_\beta \cap I^{\mathbf{u}}_\beta$ it follows that $\theta_\beta(\mathbf{v}) = g$ and $\theta_\beta(\mathbf{u}) = \beta|\mathcal{E}\mathbf{u}|$ with $\beta|\mathcal{E}\mathbf{u}| \geq g$.

Therefore, (3.12) is fulfilled, since

$$\theta_\beta(\mathbf{u}) - \theta_\beta(\mathbf{v}) = \beta|\mathcal{E}\mathbf{u}| - g < \beta|\mathcal{E}\mathbf{u}| - \beta|\mathcal{E}\mathbf{v}| \leq \beta|\mathcal{E}\mathbf{u} - \mathcal{E}\mathbf{v}|.$$

Finally, in $I_\beta^{\mathbf{u}} \cap I_\beta^{\mathbf{v}}$, we have that $\theta_\beta(\mathbf{u}) = \beta|\mathcal{E}\mathbf{u}|$ and $\theta_\beta(\mathbf{v}) = \beta|\mathcal{E}\mathbf{v}|$. Therefore, we have that

$$\theta_\beta(\mathbf{u}) - \theta_\beta(\mathbf{v}) = \beta|\mathcal{E}\mathbf{u}| - \beta|\mathcal{E}\mathbf{v}| \leq \beta|\mathcal{E}\mathbf{u} - \mathcal{E}\mathbf{v}|.$$

Thus, since the four given sets provide a disjoint partitioning of Ω, inequality (3.12) is satisfied for almost all x in Ω. □

Remark 3.1. *Let $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ be given. Then, we have that*

$$\frac{|\mathcal{E}\mathbf{u}|}{\theta_\beta(\mathcal{E}\mathbf{u})} \leq \frac{1}{\beta}, \ \textit{a.e. in } \Omega. \tag{3.14}$$

Indeed, since $\theta_\beta > 0$ and by recalling the sets $E_\beta^{\mathbf{u}}$ and $I_\beta^{\mathbf{u}}$ from Lemma 3.1, we analogously observe that the following pointwise estimates hold. On $E_\beta^{\mathbf{u}}$, we have that $\theta_\beta(\mathcal{E}\mathbf{u}) = g$ and that $|\mathcal{E}\mathbf{u}| < \frac{g}{\beta}$. Then, we obtain that

$$\frac{|\mathcal{E}\mathbf{u}|}{\theta_\beta(\mathcal{E}\mathbf{u})} = \frac{|\mathcal{E}\mathbf{u}|}{g} < \frac{1}{\beta}, \ \textit{a.e. in } \Omega.$$

On $I_\beta^{\mathbf{u}}$, we have that $\theta_\beta(\mathcal{E}\mathbf{u}) = \beta|\mathcal{E}\mathbf{u}|$. Then, we obtain that

$$\frac{|\mathcal{E}\mathbf{u}|}{\theta_\beta(\mathcal{E}\mathbf{u})} = \frac{|\mathcal{E}\mathbf{u}|}{\beta|\mathcal{E}\mathbf{u}|} = \frac{1}{\beta}, \ \textit{a.e. in } \Omega.$$

3.4 Exact Penalization Formulation

We have rewritten the Huber regularized Bingham problem as the constrained optimization problem (CP) to characterize its solutions via a KKT system. Now, the idea behind the penalty approach is to consider the constraint as a penalization of the objective functional. This approach leads us again to an unconstrained problem to be analyzed. In the case of the steady-state bi-viscous flow, taking into account the sparsification property of the L^1–norm we propose an exact penalization as follows.

$$J_\sigma(\mathbf{u}) := J(\mathbf{u}) + \sigma\|\operatorname{div}(\mathbf{u})\|_{L^1}, \tag{3.15}$$

where $\sigma > 0$ and $J(\mathbf{u})$ is given in (CP). Moreover, the functional given in (3.15) is also proper, continuous, strictly convex, and coercive, i.e.,

$$\lim_{\|\mathbf{u}\|_{\mathbf{H}_0^1} \to \infty} J_\sigma(\mathbf{u}) = +\infty.$$

Thus, from Theorem 2.3, we conclude that the minimization problem

$$\min_{\mathbf{u}\in\mathbf{H}_0^1(\Omega)} J_\sigma(\mathbf{u}) \tag{EP}$$

has a unique solution $\bar{\mathbf{u}}_\sigma \in \mathbf{H}_0^1(\Omega)$. Additionally, by Fermat's Rule (see Theorem 2.4), the optimality condition reads as follows

$$0 \in \partial J_\sigma(\bar{\mathbf{u}}_\sigma) := J'(\bar{\mathbf{u}}_\sigma) + \partial h(\bar{\mathbf{u}}_\sigma), \tag{3.16}$$

where $h(\bar{\mathbf{u}}_\sigma) = \sigma\|\operatorname{div}(\bar{\mathbf{u}}_\sigma)\|_{L^1}$ and $\partial h(\bar{\mathbf{u}}_\sigma)$ is the convex subdifferential of h at $\bar{\mathbf{u}}_\sigma$. Moreover, taking into account that $h = \sigma\|\cdot\|_{L^1} \circ \operatorname{div}$, and by using the rules of subdifferential calculus for the composition of functions (see [33, Th. 4.13]), we have that for $\eta \in \partial h(\mathbf{u})$, there exists $\zeta \in \sigma\partial\|\cdot\|_{L^1}(\operatorname{div}\mathbf{u})$, such that

$$\langle \eta, \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = \langle -\operatorname{grad}\zeta, \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = (\zeta, \operatorname{div}\mathbf{v})_{L^2},\ \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{3.17}$$

Therefore, the optimality condition (3.16) turns into

$$\langle -J'(\bar{\mathbf{u}}_\sigma), \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = (\zeta, \operatorname{div}\mathbf{v})_{L^2},\ \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{3.18}$$

Remark 3.2. *Note that $\zeta \in \sigma\partial\|\cdot\|_{L^1}(\operatorname{div}\bar{\mathbf{u}}_\sigma)$ yields that $\zeta \in L_0^2(\Omega)$, and $|\zeta| \leq \sigma$ a.e. in Ω, i.e., $\zeta \in L^\infty(\Omega)$.*

Let us now discuss about the equivalence of the constrained problem (CP) and the penalized problem (EP).

Theorem 3.1. *Let $\tilde{\mathbf{u}}$ be the solution of problem (CP). Then $\tilde{\mathbf{u}}$ is also the solution of (EP). Furthermore, let $\bar{\mathbf{u}}_\sigma$ be the solution of problem (EP), associated to a given σ. Then, there exists $\sigma_0 > 0$ such that for all $\sigma > \sigma_0$ the divergence free condition*

$$\|\operatorname{div}\bar{\mathbf{u}}_\sigma\|_{L^1} = 0,$$

holds. This fact implies that $\bar{\mathbf{u}}_\sigma$ is solution of the constrained problem (CP), i.e., $\bar{\mathbf{u}}_\sigma = \tilde{\mathbf{u}}$, for $\sigma > \sigma_0$.

Proof. Since $\operatorname{div}\tilde{\mathbf{u}} = 0$ and $J(\tilde{\mathbf{u}}) \leq J(\mathbf{u})$, it follows that

$$J(\tilde{\mathbf{u}}) + \sigma\|\operatorname{div}\tilde{\mathbf{u}}\|_{L^1} = J(\tilde{\mathbf{u}}) \leq J(\mathbf{u}) \leq J(\mathbf{u}) + \sigma\|\operatorname{div}\mathbf{u}\|_{L^1},\ \forall\, \mathbf{u} \in \mathbf{H}_0^1(\Omega).$$

Thus, the solution $\tilde{\mathbf{u}}$ of the constrained problem (CP) is also the minimizer of the functional in (EP). Moreover, applying the optimality condition (3.16) to the solution $\tilde{\mathbf{u}}$, we obtain that $-J'(\tilde{\mathbf{u}}) \in \sigma\partial\|\operatorname{div}\tilde{\mathbf{u}}\|_{L^1}$. Therefore, there exists $\zeta \in \sigma\partial\|\cdot\|_{L^1}(\operatorname{div}\tilde{\mathbf{u}})$,

such that

$$\langle -J'(\tilde{\mathbf{u}}), \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = (\zeta, \operatorname{div}\, \mathbf{v})_{L^2},\ \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{3.19}$$

By combining (3.19) with the KKT condition (3.10b), we get that

$$\langle -J'(\tilde{\mathbf{u}}), \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = (\zeta, \operatorname{div}\, \mathbf{v})_{L^2} = (-\lambda, \operatorname{div}\, \mathbf{v})_{L^2},\ \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{3.20}$$

Moreover, since $-\operatorname{div} : \mathbf{H}_0^1(\Omega) \to L_0^2(\Omega)$ is the dual operator of $\operatorname{grad} : L_0^2(\Omega) \to \mathbf{H}^{-1}(\Omega)$, from (3.20) we obtain

$$(\zeta + \lambda, \operatorname{div}\, \mathbf{v})_{L^2} = \langle -\operatorname{grad}(\zeta + \lambda), \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = 0,\ \ \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega).$$

Since $\mathbf{v}$ is arbitrary, we have that $-\operatorname{grad}(\zeta + \lambda) = 0$ in $\mathbf{H}^{-1}(\Omega)$. By the definition of the operator grad (see [31, Th. 6.14-1, p. 396]), we deduce that $(\zeta + \lambda)$ is constant almost everywhere in Ω. However, since $(\zeta + \lambda) \in L_0^2(\Omega)$, we have that $(\zeta + \lambda) = 0$. Hence, from Remark 3.2 we conclude that $\zeta = -\lambda \in L^\infty(\Omega)$.

Next, we prove the reciprocal by contradiction. Therefore let us assume that for all $\sigma_0 > 0$, there exists $\sigma > \sigma_0$ such that $\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1} > 0$. Next, since $\tilde{\mathbf{u}}$ is the solution of problem (CP), we have that $\operatorname{div} \tilde{\mathbf{u}} = 0$. Moreover, we know that $\bar{\mathbf{u}}_\sigma$ minimizes J_σ, which yields that

$$\begin{aligned} 0 &\leq J_\sigma(\tilde{\mathbf{u}}) - J_\sigma(\bar{\mathbf{u}}_\sigma) \\ &= J(\tilde{\mathbf{u}}) + \sigma\|\operatorname{div}\, \tilde{\mathbf{u}}\|_{L^1} - J(\bar{\mathbf{u}}_\sigma) - \sigma\|\operatorname{div}\, \bar{\mathbf{u}}_\sigma\|_{L^1} \\ &= J(\tilde{\mathbf{u}}) - J(\bar{\mathbf{u}}_\sigma) - \sigma\|\operatorname{div}\, \bar{\mathbf{u}}_\sigma\|_{L^1}. \end{aligned} \tag{3.21}$$

Using the fact that J is convex and differentiable, (3.21) implies that

$$\sigma\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1} \leq J(\tilde{\mathbf{u}}) - J(\bar{\mathbf{u}}_\sigma) \leq -\left\langle J'(\tilde{\mathbf{u}}), \bar{\mathbf{u}}_\sigma - \tilde{\mathbf{u}}\right\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}.$$

Then, from the optimality condition (3.10b) for $\tilde{\mathbf{u}}$, we obtain that

$$\begin{aligned} \sigma\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1} &\leq -\left\langle J'(\tilde{\mathbf{u}}), \bar{\mathbf{u}}_\sigma - \tilde{\mathbf{u}}\right\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} \\ &= -(\lambda, \operatorname{div}(\bar{\mathbf{u}}_\sigma - \tilde{\mathbf{u}}))_{L^2} \\ &\leq |(\lambda, \operatorname{div} \bar{\mathbf{u}}_\sigma)_{L^2}|. \end{aligned} \tag{3.22}$$

From the first part of the proof, we obtained that $-\lambda \in L^\infty(\Omega)$. Thus, by applying Hölder's inequality to (3.22), we have that

$$\sigma\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1} \leq |(\lambda, \operatorname{div} \bar{\mathbf{u}}_\sigma)_{L^2}| \leq \|\lambda\|_{L^\infty}\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1},$$

which yields

$$\sigma_0 < \sigma \leq \|\lambda\|_{L^\infty}. \tag{3.23}$$

This contradicts our initial assumption by taking any $\sigma_0 > \|\lambda\|_{L^\infty}$. Thus, there exists $\sigma_0 > 0$ such that, for all $\sigma > \sigma_0$, $\|\operatorname{div} \bar{\mathbf{u}}\|_{L^1} = 0$.

The last condition imply that if σ is larger than σ_0 then $\bar{\mathbf{u}}_\sigma$ is feasible for the constrained problem (CP). Further, since $\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1} = 0$, it follows that

$$J(\bar{\mathbf{u}}_\sigma) = J(\bar{\mathbf{u}}_\sigma) + \sigma\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1} = J_\sigma(\bar{\mathbf{u}}_\sigma) \leq J_\sigma(\tilde{\mathbf{u}}) = J(\tilde{\mathbf{u}}). \tag{3.24}$$

Therefore, by the definition of $\tilde{\mathbf{u}}$ we have $J(\tilde{\mathbf{u}}) = J(\bar{\mathbf{u}}_\sigma)$, where $\bar{\mathbf{u}}_\sigma$ is a global minimum for problem (CP). □

In view of the previous result, the minimization of J_σ is called *exact penalization* formulation of (CP).

Remark 3.3. *For numerical computation purposes it is derivable a sharp estimation for σ_0, which can be used a priory in order to guarantee exact penalization. We discuss this estimation for σ_0 by using Theorem 3.1. From the embedding $L^2(\Omega) \hookrightarrow L^1(\Omega)$ we have that $\|\operatorname{div} \mathbf{u}\|_{L^1} \leq |\Omega|^{\frac{1}{2}}\|\operatorname{div} \mathbf{u}\|_{L^2}$. Multiplying this inequality by $\|\lambda\|_{L^2}$, we have that*

$$\|\lambda\|_{L^2}|\Omega|^{-\frac{1}{2}}\|\operatorname{div} \mathbf{u}\|_{L^1} \leq \|\lambda\|_{L^2}\|\operatorname{div} \mathbf{u}\|_{L^2}. \tag{3.25}$$

From the proof of Theorem 3.1 it is clear that, if $\bar{\mathbf{u}}_\sigma$ is the solution of the unconstrained problem (EP) associated to a $\sigma \leq \sigma_0$, then inequality (3.22) is satisfied in a nontrivial manner, i.e., in particular

$$\sigma_0\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^1} \leq \|\lambda\|_{L^2}\|\operatorname{div} \bar{\mathbf{u}}_\sigma\|_{L^2}.$$

Therefore, from (3.25) applied to $\bar{\mathbf{u}}_\sigma$, we might consider either $\sigma_0 \geq \|\lambda\|_{L^2}|\Omega|^{-\frac{1}{2}}$ or $\sigma_0 < \|\lambda\|_{L^2}|\Omega|^{-\frac{1}{2}}$. If the later holds, we have found a $\bar{\sigma} = \|\lambda\|_{L^2}|\Omega|^{-\frac{1}{2}} > \sigma_0$. Then thanks to Theorem 3.1 we have that $\operatorname{div} \bar{\mathbf{u}}_\sigma = 0$ and equation (3.22) is satisfied trivially. On the other hand, if $\sigma_0 \geq \|\lambda\|_{L^2}|\Omega|^{-\frac{1}{2}}$, we arrive to a lower bound for σ_0. Then, with both results we can establish the estimation

$$\sigma_0 \approx \|\lambda\|_{L^2}|\Omega|^{-\frac{1}{2}}. \tag{3.26}$$

In practice, by solving the system $\langle J'(\bar{\mathbf{u}}_\sigma), \mathbf{u}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = (\lambda, \operatorname{div}(\mathbf{u}))_{L^2}$, for all $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ we can obtain λ in order to estimate σ_0.

3.5 Quadratic Penalization

In this section we discuss the quadratic penalty approach using the L^2–norm. Our aim is to show the differences and similarities with the exact penalization approach for

the Bingham viscoplastic flow problem. The quadratic penalty function, involving the L^2–norm is given by

$$J_\nu(\mathbf{u}) := J(\mathbf{u}) + \frac{\nu}{2}\int_\Omega |\operatorname{div}\mathbf{u}|^2\, dx, \tag{QP}$$

where $\nu > 0$ and $J(\mathbf{u})$ is given in (CP). Similarly to the exact penalty approach, as discussed in Section 3.4, the function in problem (QP) is also proper, continuous, and strictly convex, which satisfies

$$\lim_{\|\mathbf{u}\|_{\mathbf{H}_0^1}\to\infty} J_\nu(\mathbf{u}) = +\infty.$$

Consequently, from Theorem 2.3, we conclude that the following minimization problem has a unique solution in $\mathbf{H}_0^1$, for each $\nu > 0$.

$$\min_{\mathbf{u}\in\mathbf{H}_0^1(\Omega)} J_\nu(\mathbf{u}). \tag{3.27}$$

Let us note by $\mathbf{u}_\nu$ the solution to problem (3.27). Further, let us recall that this function must be the solution of the following PDE, which corresponds to the Euler equation associated to the optimization problem. Therefore, $\mathbf{u}_\nu$ satisfies

$$\begin{aligned} 2\mu\int_\Omega \mathcal{E}\mathbf{u}_\nu : \mathcal{E}\mathbf{v}\, dx + g\beta\int_\Omega \frac{\mathcal{E}\mathbf{u}_\nu : \mathcal{E}\mathbf{v}}{\max(g, \beta|\mathcal{E}\mathbf{u}_\nu|)}\, dx + \nu\int_\Omega (\operatorname{div}\mathbf{u}_\nu)(\operatorname{div}\mathbf{v})\, dx \\ = \int_\Omega \mathbf{f_b}\cdot\mathbf{v}\, dx,\ \forall\mathbf{v}\in\mathbf{H}_0^1(\Omega). \end{aligned} \tag{3.28}$$

3.6 Recovering the fluid's pressure

De Rham's theorem [109] is a fundamental result in differential geometry that establishes a correspondence between differential forms and cohomology classes. In fluid mechanics, this theorem is often applied in the context of incompressible flows to decompose a vector field. This decomposition is crucial for identifying the pressure field in problems where the velocity field satisfies the incompressibility condition ($\operatorname{div}\mathbf{u} = 0$). In what follows we present a coarse version of De Rham's theorem established in [55].

Theorem 3.2. *If* $\mathbf{f}\in\mathbf{H}^{-1}(\Omega)$ *satisfies*

$$\langle\mathbf{f},\mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = 0 \quad \forall\mathbf{v}\in V,$$

then there exists $p_r\in L^2(\Omega)$ *such that*

$$\mathbf{f} = \boldsymbol{grad}\ p_r.$$

Proof. See [55, Lemma 2.1]. □

Let us recall that the solution of the Huber regularized Bingham problem (CP) is also a solution to the following PDE

$$\begin{aligned} 2\mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v}\, dx + g\beta \int_\Omega \frac{\mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v}}{\max(g, \beta|\mathcal{E}\mathbf{u}|)}\, dx - \int_\Omega p_r \operatorname{div}\mathbf{v}\, dx \\ = \int_\Omega \mathbf{f_b} \cdot \mathbf{v}\, dx, \ \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \end{aligned} \tag{3.29}$$

Thus, it is well established, by De Rham's theorem, the existence of the function $p_r \in L^2(\Omega)$ that solves (3.29). This function can be seen as a Lagrange multiplier associated to the restriction $\operatorname{div}\mathbf{u} = 0$. For more details, we refer the reader to [37].

In what follows we will argue that in the case of the exact and quadratic penalty methods presented in Sections 3.4 and 3.5, there exists a function playing the role for the pressure of the fluid. This existence result is reached trough a convergence argument in the divergence of the velocity's vector field that we analyze for both penalizations. Naturally, in the case of exact penalization, the associated analysis is more challenging due to the nondiferentaility of the L^1–norm and the presence of its associated subgradients.

From the mechanical point of view, (3.29) represents the flow of an incompressible Huber regularized Bingham flow, while (3.28) represents the flow of a slightly incompressible Bingham flow. This means that $\operatorname{div}\mathbf{u}_\nu \neq 0$, but nearly to zero. In fact, we expect that $\operatorname{div}\mathbf{u}_\nu \to 0$, as $\nu \to \infty$ (see [109]). This last convergence property is recognizable different from the expected behavior of the exact penalization presented in Section 3.4, where the incompressibility of the fluid is expected to hold for a (possibly large) finite value of the penalization parameter σ.

In the following result, we first prove that the sequence generated by the quadratic penalization converges to the solution of the PDE given in (3.29).

Theorem 3.3. *Let $\{\mathbf{u}_\nu\} \subset \mathbf{H}_0^1(\Omega)$ be the sequence formed by the solutions of (3.28) associated to the parameter $\nu > 0$. Moreover, let $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ be the solution of the variational problem (3.29). Then,*

$$\mathbf{u}_\nu \to \mathbf{u}, \ \textit{in } \mathbf{H}_0^1(\Omega), \ \textit{as } \nu \to \infty. \tag{3.30}$$

Proof. By subtracting the equation (3.29) from equation (3.28), we have that the variational problem holds

$$\begin{aligned} 2\mu \int_\Omega \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) : \mathcal{E}\mathbf{v}\, dx + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\nu}{\theta_\beta(\mathbf{u}_\nu)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}\mathbf{v}\, dx \\ + \nu \int_\Omega (\operatorname{div}\mathbf{u}_\nu)(\operatorname{div}\mathbf{v})\, dx = -\int_\Omega p_r \operatorname{div}\mathbf{v}\, dx, \ \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega), \end{aligned} \tag{3.31}$$

where $\theta_\beta(\mathbf{u})$ was introduced in Lemma 3.1. Next, we take $\mathbf{v} = \mathbf{u}_\nu - \mathbf{u}$ in the above equation, where $\mathbf{u}$ fulfills the divergence–free condition $\operatorname{div} \mathbf{u} = 0$. Then, it follows that

$$\begin{aligned} 2\mu \int_\Omega |\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})|^2 \, dx + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\nu}{\theta_\beta(\mathbf{u}_\nu)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) \, dx \\ + \nu \int_\Omega |\operatorname{div} \mathbf{u}_\nu|^2 \, dx = - \int_\Omega p_r \operatorname{div} \mathbf{u}_\nu \, dx. \end{aligned} \tag{3.32}$$

Let us focus on the second term on the left hand side of (3.32). Here, Lemma 3.1 and the Cauchy-Schwarz inequality imply that

$$\begin{aligned} &\int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\nu}{\theta_\beta(\mathbf{u}_\nu)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) \, dx \\ &\qquad = \int_\Omega \left[\mathcal{E}\mathbf{u} \left(\frac{1}{\theta_\beta(\mathbf{u}_\nu)} - \frac{1}{\theta_\beta(\mathbf{u})} \right) + \frac{\mathcal{E}\mathbf{u}_\nu - \mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u}_\nu)} \right] : \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) \, dx \\ &\qquad = \int_\Omega \left[\frac{|\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})|^2}{\theta_\beta(\mathbf{u}_\nu)} + \left(\frac{\theta_\beta(\mathbf{u}) - \theta_\beta(\mathbf{u}_\nu)}{\theta_\beta(\mathbf{u}_\nu)\theta_\beta(\mathbf{u})} \right) \mathcal{E}\mathbf{u} : \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) \right] dx \\ &\qquad = \int_\Omega \frac{1}{\theta_\beta(\mathbf{u}_\nu)} \left[|\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})|^2 - \frac{\theta_\beta(\mathbf{u}_\nu) - \theta_\beta(\mathbf{u})}{\theta_\beta(\mathbf{u})} \mathcal{E}\mathbf{u} : \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) \right] dx \\ &\qquad \geq \int_\Omega \frac{1}{\theta_\beta(\mathbf{u}_\nu)} \left[|\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})|^2 - \beta \frac{|\mathcal{E}\mathbf{u}_\nu - \mathcal{E}\mathbf{u}|}{\theta_\beta(\mathbf{u})} |\mathcal{E}\mathbf{u}| |\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})| \right] dx. \end{aligned}$$

Thus, by taking into account (3.14), it holds that

$$\begin{aligned} &\int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\nu}{\theta_\beta(\mathbf{u}_\nu)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) dx \\ &\qquad \geq \int_\Omega \frac{1}{\theta_\beta(\mathbf{u}_\nu)} \left[|\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})|^2 - \beta |\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})|^2 \frac{|\mathcal{E}\mathbf{u}|}{\theta_\beta(\mathbf{u})} \right] dx \geq 0. \end{aligned}$$

Next, inserting the last relation in (3.32), and using Korn's inequality, we conclude the existence of a constant $C > 0$ such that

$$\begin{aligned} C\|\mathbf{u}_\nu - \mathbf{u}\|^2_{\mathbf{H}^1_0} + \nu \int_\Omega |\operatorname{div} \mathbf{u}_\nu|^2 \, dx \leq 2\mu \int_\Omega |\mathcal{E}(\mathbf{u}_\nu - \mathbf{u})|^2 \, dx \\ + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\nu}{\theta_\beta(\mathbf{u}_\nu)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) \, dx + \nu \int_\Omega |\operatorname{div} \mathbf{u}_\nu|^2 \, dx \\ = - \int_\Omega p_r \operatorname{div} \mathbf{u}_\nu \, dx \leq \|p_r\|_{L^2} \| \operatorname{div} \mathbf{u}_\nu\|_{L^2}. \end{aligned} \tag{3.33}$$

Further, Young's inequality implies that

$$\|p_r\|_{L^2} \| \operatorname{div} \mathbf{u}_\nu\|_{L^2} \leq \frac{\nu}{2} \| \operatorname{div} \mathbf{u}_\nu\|^2_{L^2} + \frac{1}{2\nu} \|p_r\|^2_{L^2},$$

for $\nu > 0$. By using this inequality in (3.33), we obtain that

$$C\|\mathbf{u}_\nu - \mathbf{u}\|^2_{\mathbf{H}^1_0} + \nu \int_\Omega |\operatorname{div} \mathbf{u}_\nu|^2 \, dx \leq \frac{\nu}{2} \int_\Omega |\operatorname{div} \mathbf{u}_\nu|^2 \, dx + \frac{1}{2\nu} \|p_r\|_{L^2},$$

which yields that

$$C\|\mathbf{u}_\nu - \mathbf{u}\|^2_{\mathbf{H}^1_0} + \frac{\nu}{2} \int_\Omega |\operatorname{div} \mathbf{u}_\nu|^2 \, dx \leq \frac{1}{2\nu} \|p_r\|_{L^2}.$$

Consequently,

$$C\|\mathbf{u}_\nu - \mathbf{u}\|^2_{\mathbf{H}^1_0} \leq \frac{1}{2\nu} \|p_r\|_{L^2}.$$

Taking the limit $\nu \to \infty$, we get that $\mathbf{u}_\nu \to \mathbf{u} \in \mathbf{H}^1_0(\Omega)$. □

3.6.1 Pressure in the Quadratic Penalization

In order to recover the fluid's pressure, we require the following result.

Lemma 3.2. *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$. Then there exists a constant $c = c(\Omega)$ depending only on Ω, such that, for every $w \in L^2(\Omega)$ the following result holds:*

$$\|w\|_{L^2(\Omega)} \leq c(\Omega) \left\{ \left| \int_\Omega w \, dx \right| + \sum_{i=1}^n \left\| \frac{\partial w}{\partial x_i} \right\|_{H^{-1}(\Omega)} \right\} \tag{3.34}$$

Proof. See [109, Lem. 6.1, pp. 100]. □

Next, to analyze the pressure in the quadratic penalization, we consider the difference between the incompressible regularized Bingham flow, given in equation (3.29), and the flow associated to the quadratic penalized problem given in (3.28). This difference was defined in (3.31). Accordingly, we rewrite it as follows:

$$\begin{aligned} 2\mu \int_\Omega \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) : \mathcal{E}\mathbf{v} \, dx + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\nu}{\theta_\beta(\mathbf{u}_\nu)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}\mathbf{v} \, dx \\ + \int_\Omega (\nu \operatorname{div} \mathbf{u}_\nu + p_r)(\operatorname{div} \mathbf{v}) \, dx = 0, \quad \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega). \end{aligned}$$

Given that the divergence operator is the dual operator of the gradient (see [31, Th. 6.14-1]), it holds that

$$\int_\Omega (\nu \operatorname{div} \mathbf{u}_\nu + p_r)(\operatorname{div} \mathbf{v}) \, dx = \langle \nabla(\nu \operatorname{div} \mathbf{u}_\nu + p_r), \mathbf{v} \rangle_{\mathbf{H}^{-1}, \mathbf{H}^1_0} \quad \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega).$$

Then, it follows that

$$\begin{aligned} 2\mu \int_\Omega \mathcal{E}(\mathbf{u}_\nu - \mathbf{u}) : \mathcal{E}\mathbf{v}\, dx + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\nu}{\theta_\beta(\mathbf{u}_\nu)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}\mathbf{v}\, dx \\ -\langle \nabla(\nu \operatorname{div} \mathbf{u}_\nu + p_r)\,,\, \mathbf{v} \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = 0, \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \end{aligned} \tag{3.35}$$

Next, by the continuity of θ_β and Theorem 3.3, we take the limit $\nu \to \infty$ in (3.35) and obtain that

$$\frac{\partial}{\partial x_i} (\nu \operatorname{div} \mathbf{u}_\nu) \to -\frac{\partial p_r}{\partial x_i}, \quad \text{in } \mathbf{H}^{-1}(\Omega) \text{ and for all } i = 1, \dots, n. \tag{3.36}$$

On the other hand, since $p_r \in L_0^2(\Omega)$, $\mathbf{u}_\nu = 0$ on $\partial\Omega$, and thanks to the divergence theorem, we have that

$$\begin{aligned} \int_\Omega (p_r + \nu \operatorname{div} \mathbf{u}_\nu)\, dx &= \int_\Omega p_r\, dx + \nu \int_\Omega \operatorname{div} \mathbf{u}_\nu\, dx \\ &= \int_{\partial\Omega} \mathbf{u}_\nu \cdot \vec{\mathbf{n}}\, dx = 0. \end{aligned}$$

Therefore, Lemma 3.2 yields that

$$\|p_r + \nu \operatorname{div} \mathbf{u}_\nu\|_{L^2} \leq c(\Omega) \sum_{i=1}^{n} \left\| \frac{\partial}{\partial x_i} (p_r + \nu \operatorname{div} \mathbf{u}_\nu) \right\|_{\mathbf{H}^{-1}(\Omega)},$$

which, thanks to (3.36), implies that

$$-\nu \operatorname{div} \mathbf{u}_\nu \to p_r, \text{ as } \nu \to \infty.$$

3.6.2 Pressure in the Exact Penalization

Now, we turn our discussion to the pressure recovery in the context of exact penalization. We start by recalling that the optimality condition (3.18) implies that the solution $\mathbf{u}_\sigma$ of the exact penalized problem (EP), is characterized by the existence of a function $\zeta \in L_0^2(\Omega)$, such that $|\zeta| \leq \sigma$ a.e. in Ω, satisfying:

$$\langle -J'(\mathbf{u}_\sigma), \mathbf{v} \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = (\zeta, \operatorname{div} \mathbf{v})_{L^2}\,, \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega).$$

By using the Fréchet derivative of J given in (3.11), we can observe that this equation is equivalent to the following PDE:

$$\begin{aligned} 2\mu \int_\Omega \mathcal{E}\mathbf{u}_\sigma : \mathcal{E}\mathbf{v}\, dx + g\beta \int_\Omega \frac{\mathcal{E}\mathbf{u}_\sigma : \mathcal{E}\mathbf{v}}{\theta_\beta(\mathbf{u}_\sigma)}\, dx - \int_\Omega \mathbf{f_b} \cdot \mathbf{v}\, dx \\ = -\int_\Omega \zeta \operatorname{div} \mathbf{v}\, dx, \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \end{aligned} \tag{3.37}$$

Similarly as in the quadratic case, we subtract equation (3.29) from the PDE associated to the exact penalization given in (3.37). Then we have that

$$\begin{aligned} 2\mu \int_\Omega \mathcal{E}(\mathbf{u}_\sigma - \mathbf{u}) : \mathcal{E}\mathbf{v}\, dx + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\sigma}{\theta_\beta(\mathbf{u}_\sigma)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}\mathbf{v}\, dx \\ = -\int_\Omega (\zeta + p_r) \operatorname{div} \mathbf{v}\, dx,\ \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \end{aligned} \tag{3.38}$$

Taking $\mathbf{v} = \mathbf{u}_\sigma - \mathbf{u}$ in (3.38), we obtain that

$$\begin{aligned} 2\mu \int_\Omega |\mathcal{E}(\mathbf{u}_\sigma - \mathbf{u})|^2\, dx + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\sigma}{\theta_\beta(\mathbf{u}_\sigma)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}(\mathbf{u}_\sigma - \mathbf{u})\, dx \\ = -\int_\Omega (\zeta + p_r) \operatorname{div}(\mathbf{u}_\sigma - \mathbf{u})\, dx. \end{aligned}$$

We already proved that the second term in the left hand side is non negative. Therefore, this last equation, together with Korn's inequality and the incompressibility of $\mathbf{u}$ imply the existence of a constant $C > 0$, such that

$$C\|\mathbf{u}_\sigma - \mathbf{u}\|^2_{\mathbf{H}_0^1} \leq -\int_\Omega (p_r + \zeta) \operatorname{div} \mathbf{u}_\sigma\, dx. \tag{3.39}$$

On the other hand, Theorem 3.1 states that there exists $\sigma_0 > 0$ such that $\operatorname{div} \mathbf{u}_\sigma = 0$, for all $\sigma \geq \sigma_0$. Therefore, (3.39) yields that

$$\mathbf{u}_\sigma = \mathbf{u}, \quad \text{in } \mathbf{H}_0^1(\Omega), \text{ for all } \sigma \geq \sigma_0. \tag{3.40}$$

Also, equation (3.39) and the Korn's inequality imply that

$$\int_\Omega |\mathcal{E}\mathbf{u}_\sigma - \mathcal{E}\mathbf{u}|^2\, dx = 0, \quad \text{for all } \sigma > \sigma_0,$$

which yields that $|\mathcal{E}\mathbf{u}_\sigma - \mathcal{E}\mathbf{u}|^2 = 0$, a.e. in Ω (see[54, Prop. 2.10]). Consequently, we can infer that

$$\mathcal{E}\mathbf{u}_\sigma = \mathcal{E}\mathbf{u}, \text{ a.e. in } \Omega \text{ and for all } \sigma > \sigma_0. \tag{3.41}$$

Next, we establish pointwise bounds for $\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u})$ on the next four disjoint sets: $E_\beta^{\mathbf{u}_\sigma} \cap E_\beta^{\mathbf{u}}$, $E_\beta^{\mathbf{u}_\sigma} \cap I_\beta^{\mathbf{u}}$, $E_\beta^{\mathbf{u}} \cap I_\beta^{\mathbf{u}_\sigma}$ and $I_\beta^{\mathbf{u}_\sigma} \cap I_\beta^{\mathbf{u}}$. Note that these sets were defined in Lemma 3.1.

In $E_\beta^{\mathbf{u}_\sigma} \cap E_\beta^{\mathbf{u}}$, we directly have that $\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) = 0$. In $E_\beta^{\mathbf{u}_\sigma} \cap I_\beta^{\mathbf{u}}$, we have that $\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) = g - \beta|\mathcal{E}\mathbf{u}| \leq 0$. On the other hand, thanks to (3.41), we have that $g - \beta|\mathcal{E}\mathbf{u}| = g - \beta|\mathcal{E}\mathbf{u}_\sigma| \geq 0$. Thus, we have that

$$\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) = 0, \text{ a.e. in } E_\beta^{\mathbf{u}_\sigma} \cap I_\beta^{\mathbf{u}} \text{ and for all } \sigma > \sigma_0.$$

In $E_\beta^{\mathbf{u}} \cap I_\beta^{\mathbf{u}_\sigma}$, we have that $\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) = \beta|\mathcal{E}\mathbf{u}_\sigma| - g \geq 0$. Finally, in $I_\beta^{\mathbf{u}_\sigma} \cap I_\beta^{\mathbf{u}}$, we have that $\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) = \beta|\mathcal{E}\mathbf{u}_\sigma| - \beta|\mathcal{E}\mathbf{u}|$. This expression, together with (3.41), implies that

$$\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) = 0, \text{ a.e. in } I_\beta^{\mathbf{u}_\sigma} \cap I_\beta^{\mathbf{u}} \text{ and for all } \sigma > \sigma_0.$$

Summarizing, since the four given sets provide a disjoint partitioning of Ω, we can conclude that $\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) \geq 0$, a.e. in Ω and for all $\sigma > \sigma_0$. Now, Lemma 3.1 implies that $\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) < |\mathcal{E}\mathbf{u}_\sigma - \mathcal{E}\mathbf{u}| = 0$, a.e. in Ω and for all $\sigma > \sigma_0$. Therefore, we can state that

$$\theta_\beta(\mathbf{u}_\sigma) - \theta_\beta(\mathbf{u}) = 0, \text{ a.e. in } \Omega \text{ and for all } \sigma > \sigma_0. \tag{3.42}$$

We now turn our attention to the recovery of pressure for the flow. First, note that given that the divergence operator is the dual operator of the gradient (see [31, Th. 6.14-1]), this allows us to rewrite the equation (3.38) as follows

$$\begin{aligned} 2\mu \int_\Omega \mathcal{E}(\mathbf{u}_\sigma - \mathbf{u}) : \mathcal{E}\mathbf{v}\, dx + g\beta \int_\Omega \left(\frac{\mathcal{E}\mathbf{u}_\sigma}{\theta_\beta(\mathbf{u}_\sigma)} - \frac{\mathcal{E}\mathbf{u}}{\theta_\beta(\mathbf{u})} \right) : \mathcal{E}\mathbf{v}\, dx \\ = \langle \nabla(\zeta + p_r)\,,\, \mathbf{v} \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1},\ \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \end{aligned}$$

From this equation, (3.39) and (3.42), we conclude, for all $\sigma \geq \sigma_0$, that

$$-\frac{\partial}{\partial x_i}\zeta = \frac{\partial}{\partial x_i} p_r, \text{ in } \mathbf{H}^{-1}(\Omega), \text{ for all } i = 1, \ldots, n. \tag{3.43}$$

Finally, since $p_r, \zeta \in L_0^2(\Omega)$, we have that

$$\int_\Omega (\zeta + p_r)\, dx = 0,$$

which implies, together with Lemma 3.2 and (3.43), that

$$\|p_r + \zeta\|_{L^2} \leq c(\Omega) \sum_{i=1}^{n} \left\| \frac{\partial}{\partial x_i}(\zeta + p_r) \right\|_{\mathbf{H}^{-1}} = 0, \text{ for } \sigma \geq \sigma_0.$$

Consequently, $\zeta = -p_r$, for all $\sigma \geq \sigma_0$.

3.7 Second Order Method for the Exact Penalization Formulation

We are in place to describe a descent algorithm using second–order information for solving the nonsmooth problem generated by the exact penalty formulation. This algorithm follows ideas from the nonsmooth method designed for finite dimensions in [39] and [36]. Naturally, the extension of this method to the numerical solution of viscoplastic fluids entails several analytical and numerical challenges. One crucial step of the algorithm consists of the utilization of approximated second–order information for the computation of the descent direction. We will see that this procedure allows us to compute descent directions directly in the space V. With this feature, the algorithm solves the non-constrained problem (EP) ensuring that, at each iteration k, the descent direction $\mathbf{w}_k$ satisfies $\operatorname{div} \mathbf{w}_k \approx 0$, with a prescribed precision.

3.7.1 First-order Information

Recall that the regular term $J(\mathbf{u})$ of problem (EP) involves a Huber regularization function of the Frobenius norm $\Psi(\mathcal{E}\mathbf{u})$. Moreover, let us introduce the active set:

$$E_\beta := \left\{ x \in \Omega : |\mathcal{E}\mathbf{u}(x)| < \frac{g}{\beta} \right\}.$$

In turn, the inactive set corresponds to $\Omega \backslash E_\beta$. These sets allows us writting the Fréchet derivative (3.11) as follows:

$$\begin{aligned} \langle J'(\mathbf{u}), \mathbf{v} \rangle_{\mathbf{H}^{-1}, \mathbf{H}_0^1} = 2\mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v} \, dx + g \int_{\Omega \backslash E_\beta} \frac{\mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v}}{|\mathcal{E}\mathbf{u}|} \, dx \\ + \beta \int_{E_\beta} \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v} \, dx - \int_\Omega \mathbf{f_b} \cdot \mathbf{u} \, dx. \end{aligned} \tag{3.44}$$

However, in addition to the regular part $J(\mathbf{u})$, the functional $J_\sigma(\mathbf{u})$ includes the nondifferentiable term $\sigma \| \operatorname{div}(\mathbf{u}) \|_{L^1}$. To determine a search direction, we analyze the steepest descent direction. From the characterization of the steepest descent direction for convex functions provided in the Preliminaries (see (2.12)), we have that the steepest descent direction for the function J_σ, denoted by $\bar{\boldsymbol{d}}$, is given by

$$\bar{\boldsymbol{d}} = -\arg\min_{\boldsymbol{d} \in \partial J_\sigma(\mathbf{u})} \|\boldsymbol{d}\|. \tag{3.45}$$

Note that finding the solution $\bar{\boldsymbol{d}} \in \mathbf{H}_0^1$ to problem (3.45) constitutes a constrained optimization problem. To address this, in the next section, we develop an efficient strategy for computing the search direction. This approach utilizes second-order infor-

mation to provide a descent direction without directly solving (3.45).

3.7.2 Second-order information: Generalized Differentiability and Semismoothness for Superposition Operators

Let us observe that the function $h(\mathbf{u}) = \sigma \int_\Omega |\operatorname{div} \mathbf{u}|\, dx$ is nondifferentiable if $|\operatorname{div} \mathbf{u}(x)| = 0$ for all $x \in \Omega$. Hence, in order to calculate first and second order information we will use the Huber regularization analogous to the one presented in (3.6). Thus, in this section we study the second order information available in a weak sense. This information is associated to the functional $J_\sigma(\mathbf{u}) = J(\mathbf{u}) + h(\mathbf{u})$, by using the notion of semismoothness for superposition operators. A second order information for a regularized formulation of $J(\mathbf{u})$, in finite dimension, was presented in [61].

Let us apply the Huber (local) regularization of the absolute value, introduced at the beginning of Section 3.3.1. To avoid confusion, in this section, it will be denoted as $|\cdot|_\gamma : \mathbb{R} \to \mathbb{R}$, where $\gamma > 0$ is the approximation parameter. The regularization is defined as:

$$|z|_\gamma = \begin{cases} \sigma|z| - \frac{\sigma^2}{2\gamma}, & if\ |z| \geq \frac{\sigma}{\gamma}, \\ \frac{\gamma}{2}|z|^2, & if\ |z| < \frac{\sigma}{\gamma}. \end{cases}$$

In this case, we define the div-active set as follows:

$$A_\gamma := \left\{ x \in \Omega : |\operatorname{div} \mathbf{u}(x)| < \frac{\sigma}{\gamma} \right\}, \tag{3.46}$$

and the div-inactive set as $\Omega \backslash A_\gamma$. Let us define the regularized mapping $h_\gamma(\mathbf{u}) = \sigma \int_\Omega |\operatorname{div} \mathbf{u}|_\gamma\, dx$. Analogously as in the case of J, its first Fréchet-derivative $h'_\gamma(\mathbf{u})$ is given by:

$$\begin{aligned} \left\langle h'_\gamma(\mathbf{u}), \mathbf{v} \right\rangle_{\mathbf{H}^{-1}, \mathbf{H}^1_0} &= \sigma \int_{\Omega \backslash A_\gamma} \frac{(\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v})}{|\operatorname{div} \mathbf{u}|}\, dx + \int_{A_\gamma} \gamma(\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v})\, dx \\ &= \sigma \int_\Omega \frac{\gamma(\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{v})}{\max(\sigma, \gamma|\operatorname{div} \mathbf{u}|)}\, dx, \quad \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega). \end{aligned}$$

However, neither $J(\mathbf{u})$ nor $h_\gamma(\mathbf{u})$ are twice Fréchet differentiable due to the presence of the non-differentiable max function. To obtain generalized second-order information, we employ the concept of semismoothness for superposition operators and the generalized differential framework developed in [115, Ch. 3], as reviewed in section 2.3.2. Using these concepts of generalized differentiation, we compute second–order information associated with the nondiferentiabilities in $J'(\mathbf{u})$ and $h'_\gamma(\mathbf{u})$. Thus, based on Definition (2.6) from Section 2.3.2, we analyze the semismoothness of the following Nemytskii operators in the subsequent lemmas.

Lemma 3.3. *Let $\Phi : \mathbf{H}_0^1(\Omega) \to L^1(\Omega)$ be a Nemytskii operator of the form (2.21) and defined by*

$$\Phi(\mathbf{u})(x) = \phi(F(\mathbf{u})(x) = \sigma\gamma \frac{\operatorname{div} \mathbf{u}(x)}{\max(\sigma, \gamma|\operatorname{div} \mathbf{u}(x)|)},$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is given by $\phi(a) = \sigma\gamma \frac{a}{\max(\sigma,\gamma|a|)}$ and $F : \mathbf{H}_0^1(\Omega) \to L^1(\Omega)$ is defined by $F(\mathbf{u}) = \operatorname{div} \mathbf{u}$. Then, Φ is semismooth and, $G(\mathbf{u})\mathbf{w} \in \partial^\circ \Phi(\mathbf{u})\mathbf{w}$ is given by:

$$G(\mathbf{u})\mathbf{w}(x) = \begin{cases} 0 & \text{if } \gamma|\operatorname{div} \mathbf{u}(x)| \geq \sigma \\ \gamma(\operatorname{div} \mathbf{w}(x)) & \text{if } \gamma|\operatorname{div} \mathbf{u}(x)| < \sigma. \end{cases} \tag{3.47}$$

a.e on Ω.

Proof. Thanks to Theorem 2.15, Φ is semismooth on $\mathbf{H}_0^1(\Omega)$ in the sense of Definition 2.6. The rest of the proof hinges on demonstrating that conditions a) through d) from Assumption 2.1 are fulfilled, i.e., we have to prove that: $\operatorname{div} : \mathbf{H}_0^1(\Omega) \to L^1(\Omega)$ is continuously Fréchet differentiable and locally Lipschitz continuous and that $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous and semismooth. Since div is a continuous linear operator from $\mathbf{H}_0^1(\Omega)$ into $L_0^2(\Omega)$ then, it is Fréchet differentiable and its derivative is the operator div itself (see [31, Th. 6.14-1]). Additionally, the continuous embedding $L_0^2(\Omega) \subset L^1(\Omega)$ implies that the operator div is also continuously Fréchet differentiable on $L^1(\Omega)$. It is clear that div is Lipschitz continuous, hence conditions a) - b) are verified.

The proof of the Lipschitz continuity and semismoothness of ϕ is presented as follows. Let us start by rewriting $\phi(a)$ as $\phi(a) = \gamma\sigma \frac{a}{\phi_m(a)}$, with $\phi_m(a) := \max(\sigma, \gamma|a|)$. Next, we notice that the max function is globally Lipschitz continuous with constant L_{max}. This fact implies that

$$\begin{aligned} |\phi_m(a_1) - \phi_m(a_2)| &= |\max(\sigma, \gamma|a_1|) - \max(\sigma, \gamma|a_2|)| \\ &\leq \gamma L_{max} ||a_1| - |a_2|| \leq \gamma L_{max}|a_1 - a_2|, \ \forall a_1, a_2 \in \mathbb{R}. \end{aligned} \tag{3.48}$$

We conclude that ϕ_m is Lipschitz continuous. Then, we have that

$$\begin{aligned} |\phi(a_1) - \phi(a_2)| &= \left|\gamma\sigma \tfrac{a_1}{\phi_m(a_1)} - \gamma\sigma \tfrac{a_2}{\phi_m(a_2)}\right| \\ &= \gamma\sigma \left|\tfrac{a_1}{\phi_m(a_1)} - \tfrac{a_2}{\phi_m(a_2)} + \tfrac{a_1}{\phi_m(a_2)} - \tfrac{a_1}{\phi_m(a_2)}\right| \\ &\leq \gamma\sigma \left|a_1 \left(\tfrac{\phi_m(a_2) - \phi_m(a_1)}{\phi_m(a_1)\phi_m(a_2)}\right)\right| + \gamma\sigma \left|\tfrac{1}{\phi_m(a_2)}(a_1 - a_2)\right|. \end{aligned}$$

Now, it is clear that $0 < \sigma \leq \phi_m(a_2)$, which implies that $\frac{1}{\phi_m(a_2)} \leq \frac{1}{\sigma}$. By plugging this

inequality in the last expression, we have that

$$\begin{aligned} |\phi(a_1) - \phi(a_2)| &\leq \gamma\sigma \left| \frac{a_1}{\phi_m(a_1)} \left(\frac{\phi_m(a_2) - \phi_m(a_1)}{\sigma} \right) \right| + \gamma|a_1 - a_2| \\ &= \gamma \left| \frac{a_1}{\phi_m(a_1)} \right| |\phi_m(a_2) - \phi_m(a_1)| + \gamma|a_1 - a_2|. \end{aligned}$$

Finally, sice $\left| \frac{a_1}{\phi_m(a_1)} \right| \leq \frac{1}{\gamma}$, we conclude, thanks to (3.48), that

$$|\phi(a_1) - \phi(a_2)| \leq \gamma(L_{max} + 1)|a_1 - a_2| \forall a_1, a_2 \in \mathbb{R}.$$

Regarding the semismoothness of ϕ, note that the absolute value $|\cdot| : \mathbb{R} \to \mathbb{R}$ and the function $\max(0, \cdot) : \mathbb{R} \to \mathbb{R}$ are both semismooth (see [115, Sec. 2.5] and Example 2.2 respectively). Then, since the composition of semismooth functions in $\mathbb{R}^n$ is a semismooth function [115, Prop. 2.9], it follows that $\phi(a)$ is semismooth.

Then, conditions c) - d) are satisfied.

Next, we obtain a measurable selection (see Definition 2.6) $M(\mathbf{u})$ of Clarke's generalized Jacobian $\partial_C \phi(\operatorname{div} \mathbf{u})$ as follows: let $\phi_m = \phi_1 \circ \phi_2$, where $\phi_1(z) = \max(0, z) + \sigma$ and $\phi_2(y) = \gamma|y| - \sigma$. Then the following identity holds:

$$\phi_m(y) = \max(\sigma, \gamma|y|) = \max(0, \gamma|y| - \sigma) + \sigma.$$

From Example 2.2 we have that $M_{\phi_1}(\gamma|y| - \sigma) \in \partial_C \phi_1(\gamma|y| - \sigma)$, given by

$$M_{\phi_1}(\gamma|y| - \sigma) = \begin{cases} 1, & if\ \gamma|y| - \sigma > 0 \\ 0, & if\ \gamma|y| - \sigma \leq 0, \end{cases}$$

is a measurable selection of $\partial \phi_1(\gamma|y| - \sigma)$. Next, since ϕ_2 involves the function $|\cdot|$ evaluated at $y \neq 0$, from [115, Exaple 2.5.1] we have that

$$M_{\phi_2}(y) \in \partial_C \phi_2(y) = \left\{ \frac{\gamma y}{|y|} \right\} \text{for } y \neq 0.$$

Moreover, the chain rule for Clarke's generalized Jacobian [115, Prop. 2.3] yields that:

$$M_{\phi_m}(y)v \in \partial_C \phi_m(y)v \subset co\{M_{\phi_1} M_{\phi_2} v : M_{\phi_1} \in \partial_C \phi_1(\phi_2(y)), M_{\phi_2} \in \partial_C \phi_2(y)\}.$$

Thus, since $y \neq 0$,

$$M_{\phi_m}(y) = \begin{cases} \frac{\gamma y}{|y|}, & if\ \gamma|y| - \sigma > 0, \\ 0, & if\ \gamma|y| - \sigma \leq 0. \end{cases} \tag{3.49}$$

Clearly, $\phi(y) = \sigma\gamma\frac{y}{\phi_m(y)}$. Then, from the composition of functions we obtain that

$$M_\phi(y) \in \partial_C\phi(y) \subseteq \sigma\gamma\frac{\phi_m(y)\cdot 1 - y\partial_C\phi_m(y)}{\phi_m(y)^2}.$$

Then, from (3.49) the following cases can occur:

- $\gamma|y| > \sigma$. Here we have that:

$$M_\phi(y) = \sigma\frac{1}{|y|} - \sigma\frac{y^2}{|y|^3} = 0.$$

- $\gamma|y| \leq \sigma$ yields that:

$$M_\phi(y) = \gamma.$$

Finally, by taking $y = \operatorname{div}\mathbf{u}(x)$ we have the desired result

$$M(\mathbf{u}(x)) = \begin{cases} 0, & if\ \gamma|\operatorname{div}\mathbf{u}(x)| \geq \sigma \\ \gamma, & if\ \gamma|\operatorname{div}\mathbf{u}(x)| < \sigma, \end{cases} \tag{3.50}$$

a. e on Ω.

Then, Definition 2.6 and $F'(\mathbf{u})\mathbf{w} = \operatorname{div}\mathbf{w}$ yields that $G(\mathbf{u})\mathbf{w} = M(\mathbf{u})(F'(\mathbf{u})\mathbf{w}) = M(\mathbf{u})\operatorname{div}\mathbf{w}$. Finally, from (3.50) we obtain that $G(\mathbf{u})\mathbf{w} \in \partial^\circ\Phi(\mathbf{u})(\mathbf{w})$ is given by (3.47). □

In order to prove that $J'(\mathbf{u})$ is semismooth we introduce the Lemma below; the arguments of the proof are analogous to Lemma 3.3.

Lemma 3.4. *Let $\Theta : \mathbf{H}_0^1(\Omega) \to \mathbb{L}^1(\Omega)$ be a Nemytskii operator of the form (2.21) and given by*

$$\Theta\mathbf{u}(x) = g\beta\frac{\mathcal{E}\mathbf{u}(x)}{\max(g, \beta|\mathcal{E}\mathbf{u}(x)|)},$$

where Θ maps $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ to a matrix of Lebesgue functions. Then, Θ is semismooth and $K(\mathbf{u})\mathbf{w} \in \partial^\circ\Theta(\mathbf{u})\mathbf{w}$ is given by:

$$K(\mathbf{u})\mathbf{w}(x) = \begin{cases} g\dfrac{\mathcal{E}\mathbf{w}(x)}{|\mathcal{E}\mathbf{u}(x)|} - g\dfrac{(\mathcal{E}\mathbf{u}(x) : \mathcal{E}\mathbf{w}(x))\mathcal{E}\mathbf{u}(x)}{|\mathcal{E}\mathbf{u}(x)|^3}, & if\ \beta|\mathcal{E}\mathbf{u}(x)| \geq g, \\ \beta\,\mathcal{E}\mathbf{w}(x), & if\ \beta|\mathcal{E}\mathbf{u}(x)| < g. \end{cases} \tag{3.51}$$

a.e on Ω.

Proof. Let $\Theta : \mathbf{H}_0^1 \to \mathbb{L}^1(\Omega)$ be the matrix operator given by $\Theta\mathbf{u} = (\Theta_{kl}\mathbf{u})$ for $1 \leq k, l \leq n$. Further, since $\mathcal{E}\mathbf{u} = (\mathcal{E}_{kl}\mathbf{u}) \in \mathbb{L}^2(\Omega) \subset \mathbb{L}^1(\Omega)$, let us reshape this matrix as the vector

$(\mathcal{E}_j\mathbf{u})_{j=1}^m = (\mathcal{E}_1\mathbf{u}, \cdots, \mathcal{E}_m\mathbf{u})$ where $m = n^2$. Then, we have that $(\mathcal{E}_j\mathbf{u})_{j=1}^m \in \prod_{j=1}^m L^1(\Omega)$. In what follows we shall construct a Nemytskii operator, $\Theta_j : \mathbf{H}_0^1(\Omega) \to L^1(\Omega)$, of the form presented in (2.21) associated to each element $\mathcal{E}_j$ as follows:

$$\Theta_j = \varphi_j(H(\mathbf{u})) = g\beta\frac{\mathcal{E}_j\mathbf{u}(x)}{\max(g, \beta|\mathcal{E}\mathbf{u}(x)|)}, \quad \text{for } 1 \leq j \leq m.$$

Here, $H : \mathbf{H}_0^1(\Omega) \to \prod_{j=1}^m L^1(\Omega)$ maps $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ to a vector of Lebesgue functions and it is defined by $H(\mathbf{u}) = (\mathcal{E}_j\mathbf{u})_{j=1}^m$. Additionally, $\varphi_j : \mathbb{R}^m \to \mathbb{R}$ is given by $\varphi_j(y) = \frac{y_j}{\max(g,\beta|y|)}$.

Next, we will prove that each element Θ_j is semismooth by means of Theorem 2.15. Further, this argument will also prove that the operator Θ is semismooth. Analogous to the proof of Lemma 3.3, we need to verify that conditions a) - d) from Assumption 2.1 are satisfied in order to prove semismoothness of $\Theta_j(\mathbf{u})$ for $1 \leq j \leq m$. The operator $H(\mathbf{u}) = (\mathcal{E}_j(\mathbf{u}))_{j=1}^m$ is continuously Fréchet differentiable in $\prod_{j=1}^m L^2(\Omega)$. The embedding $L^2(\Omega) \subset L^1(\Omega)$ implies that $H(\mathbf{u})$ is also differentiable in $\prod_{j=1}^m L^1(\Omega)$. Moreover, linearity and boundedness of the $\mathcal{E}$ implies Lipschitz continuity of the mapping H. The Lipschitz continuity and semismoothness of φ_j, for $1 \leq j \leq m$ are obtsined by similar arguments presented for ϕ in Lemma 3.3. Then, conditions a) - d) are verified. Thus, Θ_j for $1 \leq j \leq m$ is semismooth in the sense of Definition 2.6. Now, it remains to prove that the operator $\Theta : \mathbf{H}_0^1 \to \mathbb{L}^1(\Omega)$ is also semismooth. Following [115, Prop. 3.6] let us define the set value mapping:

$$(\partial^\circ\Theta_1 \times \cdots \times \partial^\circ\Theta_m) : V \rightrightarrows \mathcal{L}(\mathbf{H}_0^1, \prod_{j=1}^m L^1),$$

where $(\partial^\circ\Theta_1 \times \cdots \times \partial^\circ\Theta_m)(\mathbf{u})$ is the set of all operators $K \in \mathcal{L}(\mathbf{H}_0^1, \prod_{j=1}^m L^1)$ of the form

$$K : \mathbf{w} \mapsto (K_1\mathbf{w}, \cdots, K_m\mathbf{w}) = (K_j\mathbf{w})_{j=1} \quad \text{with} \quad K_j \in \partial^\circ\Theta_j(\mathbf{u}),\ 1 \leq j \leq m.$$

Moreover, let us consider the nonempty set value mapping $\partial^\circ\Theta : V \rightrightarrows \mathcal{L}(\mathbf{H}_0^1, \prod_{j=1}^m L^1)$ such that $\partial^\circ\Theta(\mathbf{u}) \subset (\partial^\circ\Theta_1 \times \cdots \times \partial^\circ\Theta_m)(\mathbf{u})$ for all $\mathbf{u} \in \mathbf{H}_0^1$. Hence, $K \in \partial^\circ\Theta(\mathbf{u}+\mathbf{h})$ implies that $K \in (\partial^\circ\Theta_1 \times \cdots \times \partial^\circ\Theta_m)(\mathbf{u}+\mathbf{h})$, i.e., we obtain $K\mathbf{w} = (K_j\mathbf{w})_{j=1}^m$. Next, the space $\prod_{j=1}^m L^1$ is equipped with the norm $\|y\|_{\prod_j L^1} = \sum_{j=1}^m \|y_j\|_{L^1}$. Therefore, by the semismoothness of each Θ_j for $1 \leq j \leq m$ we have that

$$\begin{aligned}\sup_{K\in\partial^\circ\Theta(\mathbf{u}+\mathbf{h})} \|\Theta(\mathbf{u}+\mathbf{h}) - \Theta(\mathbf{u}) - K\mathbf{h}\|_{\mathbb{L}^1} &\leq \sum_{j=1}^m \sup_{K_j\in\partial^\circ\Theta_j(\mathbf{u}+\mathbf{h})} \|\Theta_j(\mathbf{u}+\mathbf{h}) - \Theta_j(\mathbf{u}) - K_j\mathbf{h}\|_{L^1} \\ &= o(\|\mathbf{h}\|_{\mathbf{H}_0^1}) \quad \text{as } \|\mathbf{h}\|_{\mathbf{H}_0^1} \to 0.\end{aligned}$$

Thus, we conclude that Θ is semismooth in the sense of Definition 2.6.

The second part of the proof consists of finding the operator $K = (K_j)_{j=1}^m \in \partial^\circ\Theta(\mathbf{u})$.

Following Definition 2.6 - equation (2.23) we have that $K_j(\mathbf{u})\mathbf{w} = N_j(\mathbf{u})^\top H'(\mathbf{u})(\mathbf{w})$, where $N_j(\mathbf{u}(x))$ (a measurable selection of Clarke's generalized Jacobian $\partial_C \varphi_j(\mathcal{E}\mathbf{u}(x))$) is given by:

$$N_j^\top(\mathbf{u}(x)) = \begin{cases} g\dfrac{e_j^\top}{|\mathcal{E}\mathbf{u}(x)|} - g\dfrac{e_j^\top \mathcal{E}\mathbf{u}(x)\mathcal{E}\mathbf{u}(x)^\top}{|\mathcal{E}\mathbf{u}(x)|^3}, & if\ \beta|\mathcal{E}\mathbf{u}(x)| \geq g, \\ \beta e_j^\top, & if\ \beta|\mathcal{E}\mathbf{u}(x)| < g, \end{cases}$$

here e_j stands for the canonical unit vector. In addition, since $H'(\mathbf{u})(\mathbf{w}) = \mathcal{E}\mathbf{w}$, we obtain that

$$K_j(\mathbf{u})\mathbf{w} = \begin{cases} g\dfrac{\mathcal{E}_j\mathbf{w}(x)}{|\mathcal{E}\mathbf{u}(x)|} - g\dfrac{\mathcal{E}_j\mathbf{u}(x)(\mathcal{E}\mathbf{u}(x)^\top \mathcal{E}\mathbf{w}(x))}{|\mathcal{E}\mathbf{u}(x)|^3}, & \text{if } \beta|\mathcal{E}\mathbf{u}(x)| \geq g, \\ \beta\, \mathcal{E}_j(\mathbf{w})(x), & \text{if } \beta|\mathcal{E}\mathbf{u}(x)| < g. \end{cases}$$

Finally, since $K = (K_j)_{j=1}^m$ and the inner product of the vectorized matrices coincides with the Frobenius product of the matrices we obtain the desired result:

$$K(\mathbf{u})\mathbf{w} = \begin{cases} g\dfrac{\mathcal{E}\mathbf{w}(x)}{|\mathcal{E}\mathbf{u}(x)|} - g\dfrac{(\mathcal{E}\mathbf{u}(x) : \mathcal{E}\mathbf{w}(x))\mathcal{E}\mathbf{u}(x)}{|\mathcal{E}\mathbf{u}(x)|^3}, & \text{if } \beta|\mathcal{E}\mathbf{u}(x)| \geq g, \\ \beta\, \mathcal{E}\mathbf{w}(x), & \text{if } \beta|\mathcal{E}\mathbf{u}(x)| < g, \end{cases}$$

a.e on Ω. □

Corollary 3.1. *The mapping* $\Theta(\mathbf{u})(x) = g\beta\dfrac{\mathcal{E}(\mathbf{u})(x)}{\max(g, \beta|\mathcal{E}\mathbf{u}(x)|)}$ *is Lipschitz continuous with Lipschitz constant* $\mathbf{L}$*, i.e.,*

$$\|\Theta(\mathbf{u_1}) - \Theta(\mathbf{u_2})\|_{\mathbb{L}^2} \leq \mathbf{L}\|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}^1}. \tag{3.52}$$

Proof. The Lipschitz continuity of Θ_j for $1 \leq j \leq m$ follows immediately from its semismoothness and by applying [115, Prop. 3.36], i.e., we have that:

$$\|\Theta_j(\mathbf{u_1}) - \Theta_j(\mathbf{u_2})\|_{L^2} \leq L_j\ \|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}_0^1}.$$

Further, since by definition of Θ we have that $\|\Theta(\mathbf{u})\|_{\mathbb{L}^2} = \left(\sum_{j=1}^m \|\Theta_j(\mathbf{u})\|_{L^2}^2\right)^{1/2}$ then, it follows that:

$$\|\Theta(\mathbf{u_1}) - \Theta(\mathbf{u_2})\|_{\mathbb{L}^2} \leq \mathbf{L}\|\mathbf{u}_1 - \mathbf{u}_2\|_{\mathbf{H}_0^1},$$

where $\mathbf{L} = \left(\sum_{j=1}^m L_j^2\right)^{1/2}$. □

Generalized Second-order Derivatives

We proceed to compute generalized second–order derivatives for $h'_\gamma(\mathbf{u})$ and $J'(\mathbf{u})$. In fact, from the semismoothness of Φ and Θ we utilize $G(\mathbf{u})\mathbf{w} \in \partial^\circ\Phi(\mathbf{u})\mathbf{w}$ and $K(\mathbf{u})\mathbf{w} \in \partial^\circ\Theta(\mathbf{u})\mathbf{w}$ to construct the generalized derivatives of $h'_\gamma(\mathbf{u})$ and $J'(\mathbf{u})$. We will denote them by $\mathcal{H}(\mathbf{u})$ and $\mathcal{J}(\mathbf{u})$ respectively. For $h'_\gamma(\mathbf{u})$ we have that

$$\left\langle h'_\gamma(\mathbf{u}), \mathbf{v} \right\rangle_{\mathbf{H}^{-1},\mathbf{H}^1_0} = \int_\Omega \operatorname{div}\mathbf{v}\ \Phi(\mathbf{u})\, dx.$$

Thus, since $G(\mathbf{u})\mathbf{w}$ is an element of the generalized differential $\partial^\circ\Phi(\mathbf{u})\mathbf{w}$, the generalized second–order derivative for $h'_\gamma(\mathbf{u})$, is given by

$$\begin{aligned} \langle \mathcal{H}(\mathbf{u})\mathbf{v}, \mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}^1_0} &= \int_\Omega \operatorname{div}\mathbf{v}\ G(\mathbf{u})\mathbf{w}\ dx \\ &= \int_{A_\gamma} \gamma \operatorname{div}\mathbf{v} \operatorname{div}\mathbf{w}\ dx, \end{aligned} \tag{3.53}$$

with A_γ defined in (3.46).

Regarding to $J'(\mathbf{u})$, given by (3.11), we note that it can be written as the sum of two differentiable terms and one nonsmooth given by

$$g\int_\Omega \beta \frac{\mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v}}{\max(g, \beta|\mathcal{E}\mathbf{u}|)}\ dx = \int_\Omega \mathcal{E}\mathbf{v} : \Theta(\mathbf{u})\ dx.$$

Next, since $K(\mathbf{u})\mathbf{w} \in \partial^\circ\Theta(\mathbf{u})\mathbf{w}$, we get

$$\begin{aligned} \langle \mathcal{J}(\mathbf{u})\mathbf{v}, \mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}^1_0} &= 2\mu\int_\Omega \mathcal{E}\mathbf{v} : \mathcal{E}\mathbf{w}\ dx + \int_\Omega K(\mathbf{u})\mathbf{w} : \mathcal{E}\mathbf{v}\ dx \\ &= 2\mu\int_\Omega \mathcal{E}\mathbf{v} : \mathcal{E}\mathbf{w}\ dx + g\int_{\Omega\setminus E_\beta} \frac{\mathcal{E}\mathbf{v} : \mathcal{E}\mathbf{w}}{|\mathcal{E}\mathbf{u}|}\ dx \\ &\quad - g\int_{\Omega\setminus E_\beta} \frac{(\mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{w})(\mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{v})}{|\mathcal{E}\mathbf{u}|^3}\ dx + \beta\int_{E_\beta} \mathcal{E}\mathbf{v} : \mathcal{E}\mathbf{w}\ dx. \end{aligned} \tag{3.54}$$

Taking advantage of the semismoothness properties of $J'(\mathbf{u}) + h'_\gamma(\mathbf{u})$, we devise a descent direction preconditioned by the second order information $\mathcal{G} = \mathcal{J}+\mathcal{H}$. Hereafter, without risk of confusion, we omit the dependence of $\mathbf{u}$ in $\mathcal{G}$. In addition, since $\mathcal{G}$ is a symmetric bilinear form, the descent direction, denoted in the same manner by $\mathbf{w} \in \mathbf{H}^1_0(\Omega)$, is obtained by solving the following system:

$$\langle \mathcal{G}\mathbf{w}, \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}^1_0} = -(\bar{\boldsymbol{d}}, \mathbf{v})_{\mathbf{H}^1_0},\ \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega),$$

or equivalently:

$$\langle(\mathcal{J}+\mathcal{H})\mathbf{w},\mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = -(\bar{\boldsymbol{d}},\mathbf{v})_{\mathbf{H}_0^1}\ \forall \mathbf{v}\in\mathbf{H}_0^1. \tag{3.55}$$

Let us recall that $\bar{\mathbf{d}}$ is given by (3.45).

In what follows, whenever there is risk of confusion, the dependence on the domain will be included in the subscript of a norm.

Lemma 3.5. *System (3.55) has a unique solution $\mathbf{w}\in\mathbf{H}_0^1(\Omega)$, which depends continuously on the descent direction $\bar{\boldsymbol{d}}\in\mathbf{H}_0^1(\Omega)$, i.e., there is a positive constant C such that*

$$\|\mathbf{w}\|_{\mathbf{H}_0^1} \leq \frac{1}{C}\|\bar{\boldsymbol{d}}\|_{\mathbf{H}_0^1}. \tag{3.56}$$

Proof. Since $\mathcal{G}\in\mathcal{L}(\mathbf{H}_0^1(\Omega),\mathbf{H}^{-1}(\Omega))$ the bilinear form $a(\cdot,\cdot):\mathbf{H}_0^1(\Omega)\times\mathbf{H}_0^1(\Omega)\to\mathbb{R}$, defined by $a(\mathbf{w},\mathbf{v})=\langle\mathcal{G}\mathbf{w},\mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}$, satisfies the hypothesis of the Babušca-Lax-Milgram Theorem [6, Th. 2.1]. Indeed, we will prove that there exist positive constants C and $\tilde{C}$ such that, for all $\mathbf{v},\mathbf{w}\in\mathbf{H}_0^1(\Omega)$, the following relations are satisfied:

(i) $a(\mathbf{w},\mathbf{w})_{\mathbf{H}_0^1}\geq C\|\mathbf{w}\|^2_{\mathbf{H}_0^1}$

(ii) $|a(\mathbf{w},\mathbf{v})_{\mathbf{H}_0^1}|\leq \tilde{C}\|\mathbf{w}\|_{\mathbf{H}_0^1}\|\mathbf{v}\|_{\mathbf{H}_0^1}$

For the first part (i), let us recall that $E_\beta := \left\{x\in\Omega : |\mathcal{E}\mathbf{u}(x)|<\frac{g}{\beta}\right\}$. Coercivity of a follows from taking $\mathbf{w}=\mathbf{v}$ in (3.54), i.e.,

$$\begin{aligned}\langle\mathcal{J}\mathbf{w},\mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = 2\mu\int_\Omega|\mathcal{E}\mathbf{w}|^2 + g\int_{\Omega\setminus E_\beta}\frac{|\mathcal{E}\mathbf{w}|^2}{|\mathcal{E}\mathbf{u}|}\,dx\\ -g\int_{\Omega\setminus E_\beta}\frac{(\mathcal{E}\mathbf{u}:\mathcal{E}\mathbf{w})^2}{|\mathcal{E}\mathbf{u}|^3}\,dx + \int_{E_\beta}\beta\,|\mathcal{E}\mathbf{w}|^2\,dx.\end{aligned}$$

By applying Cauchy-Schwarz to the Frobenius product in the third term of the right hand side, we have that

$$\begin{aligned}\langle\mathcal{J}\mathbf{w},\mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} &\geq 2\mu\|\mathcal{E}\mathbf{w}\|^2_{\mathbb{L}^2(\Omega)} + g\int_{\Omega\setminus E_\beta}\frac{|\mathcal{E}\mathbf{w}|^2}{|\mathcal{E}\mathbf{u}|}\,dx\\ &\quad -g\int_{\Omega\setminus E_\beta}\frac{|\mathcal{E}\mathbf{u}|^2|\mathcal{E}\mathbf{w}|^2}{|\mathcal{E}\mathbf{u}|^3}\,dx + \beta\|\mathcal{E}\mathbf{w}_k\|^2_{\mathbb{L}^2(E_\beta)}\\ &= 2\mu\|\mathcal{E}\mathbf{w}\|^2_{\mathbb{L}^2(\Omega)} + \beta\|\mathcal{E}\mathbf{w}\|^2_{\mathbb{L}^2(E_\beta)}.\end{aligned} \tag{3.57}$$

Further, Korn's inequality (see [78] and [120, pp. 82]) applied to $\|\mathcal{E}\mathbf{w}\|^2_{\mathbb{L}^2(\Omega)}$ implies that

$$\langle\mathcal{J}\mathbf{w},\mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}\geq C\|\mathbf{w}\|^2_{\mathbf{H}_0^1},\ \forall\,\mathbf{w}\in\mathbf{H}_0^1. \tag{3.58}$$

Similarly, taking $\mathbf{v} = \mathbf{w}$ in (3.53) yields that

$$\begin{aligned} \langle \mathcal{H}\mathbf{w}, \mathbf{w} \rangle_{\mathbf{H}^{-1}, \mathbf{H}_0^1} &= \int_{A_\gamma} \gamma (\operatorname{div} \mathbf{w})^2 \\ &= \gamma \| \operatorname{div} \mathbf{w} \|^2_{L^2(A_\gamma)}. \end{aligned} \tag{3.59}$$

This equality, together with (3.58) imply the coercivity of the bilinear form a. In fact, we have that

$$a(\mathbf{w}, \mathbf{w})_{\mathbf{H}_0^1} = \langle (\mathcal{J} + \mathcal{H})\mathbf{w}, \mathbf{w} \rangle_{\mathbf{H}^{-1}, \mathbf{H}_0^1} \geq C \|\mathbf{w}\|^2_{\mathbf{H}_0^1}, \ \forall \, \mathbf{w} \in \mathbf{H}_0^1. \tag{3.60}$$

Continuity of a follows from (3.54). Analogous to the previous arguments, we use Cauchy-Schwarz inequality resulting in

$$\begin{aligned} |\langle \mathcal{J}\mathbf{v}, \mathbf{w} \rangle_{\mathbf{H}^{-1}, \mathbf{H}_0^1}| \ &\leq \ 2\mu \|\mathcal{E}(\mathbf{v})\|_{\mathbb{L}^2(\Omega)} \|\mathcal{E}(\mathbf{w})\|_{\mathbb{L}^2(\Omega)} + g \int_{\Omega \setminus E_\beta} \frac{|\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}|}{|\mathcal{E}\mathbf{u}|} \, dx \\ &\quad + g \int_{\Omega \setminus E_\beta} \frac{|\mathcal{E}\mathbf{u}|^2 |\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}|}{|\mathcal{E}\mathbf{u}|^3} + \beta \int_{E_\beta} |\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}| \, dx \\ &= \ 2\mu \|\mathcal{E}(\mathbf{v})\|_{\mathbb{L}^2(\Omega)} \|\mathcal{E}(\mathbf{w})\|_{\mathbb{L}^2(\Omega)} + 2g \int_{\Omega \setminus E_\beta} \frac{|\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}|}{|\mathcal{E}\mathbf{u}|} \, dx \\ &+ \ \beta \int_{E_\beta} |\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}| \, dx \end{aligned} \tag{3.61}$$

Observe that in the set $\Omega \setminus E_\beta$ we have that $\frac{1}{|\mathcal{E}\mathbf{u}(x)|} \leq \frac{\beta}{g}$. Thus, in view of inequality (3.61), it follows that

$$\begin{aligned} |\langle \mathcal{J}\mathbf{v}, \mathbf{w} \rangle_{\mathbf{H}^{-1}, \mathbf{H}_0^1}| \ &\leq 2\mu \|\mathcal{E}(\mathbf{v})\|_{\mathbb{L}^2(\Omega)} \|\mathcal{E}(\mathbf{w})\|_{\mathbb{L}^2(\Omega)} \\ &\quad + 2 \int_{\Omega \setminus E_\beta} \beta |\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}| \, dx + \int_{E_\beta} \beta \, |\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}| \, dx. \\ &\leq 2\mu \|\mathcal{E}\mathbf{v}\|_{\mathbb{L}^2(\Omega)} \|\mathcal{E}\mathbf{w}\|_{\mathbb{L}^2(\Omega)} + 3 \int_\Omega \beta |\mathcal{E}\mathbf{v}||\mathcal{E}\mathbf{w}| \, dx. \end{aligned} \tag{3.62}$$

Here, we apply Hölder inequality, which results in

$$\begin{aligned} |\langle \mathcal{J}\mathbf{v}, \mathbf{w} \rangle_{\mathbf{H}^{-1}, \mathbf{H}_0^1}| \ &\leq \ 2\mu \|\mathcal{E}\mathbf{v}\|_{\mathbb{L}^2(\Omega)} \|\mathcal{E}\mathbf{w}\|_{\mathbb{L}^2(\Omega)} \\ &+ \ 3\beta \Big(\int_\Omega |\mathcal{E}\mathbf{v}|^2 \Big)^{1/2} \Big(\int_\Omega |\mathcal{E}\mathbf{w}|^2 \Big)^{1/2} \, dx \\ &= \ (2\mu + 3\beta) \|\mathcal{E}\mathbf{v}\|_{\mathbb{L}^2(\Omega)} \|\mathcal{E}\mathbf{w}\|_{\mathbb{L}^2(\Omega)}, \\ &\leq \ C_1 \|\mathbf{v}\|_{\mathbf{H}_0^1(\Omega)} \|\mathbf{w}\|_{\mathbf{H}_0^1(\Omega)}, \end{aligned} \tag{3.63}$$

where C_1 depends on μ, β and the positive constant of the continuity property of the linear operator $\mathcal{E}$. We follow a similar procedure for the term $\langle \mathcal{H}\mathbf{v}, \mathbf{w} \rangle$. By applying

Cauchy-Schwarz inequality to the inner products we estimate:

$$|\langle \mathcal{H}\mathbf{v}, \mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}| \leq \gamma \int_{A_\gamma} |\operatorname{div}\mathbf{v}||\operatorname{div}\mathbf{w}|.$$

Once again, we apply Hölder inequality and since $|\Omega| > |A_\gamma|$ we get:

$$\begin{aligned} |\langle \mathcal{H}\mathbf{v}, \mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}| &\leq \gamma \|\operatorname{div}\mathbf{v}\|_{L^2(\Omega)} \|\operatorname{div}\mathbf{w}\|_{L^2(\Omega)} \\ &\leq C_2 \|\mathbf{v}\|_{\mathbf{H}_0^1(\Omega)} \|\mathbf{w}\|_{\mathbf{H}_0^1(\Omega)} \end{aligned} \tag{3.64}$$

Here, C_2 depends on the continuity property of the divergence operator.

Finally, from the symmetry of a, equations (3.63), (3.64) and taking $\tilde{C} = \max\{C_1, C_2\}$ we have that

$$|a(\mathbf{w}, \mathbf{v})_{\mathbf{H}_0^1}| = |\langle \mathcal{G}\mathbf{w}, \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}| \leq \tilde{C}\|\mathbf{w}\|_{\mathbf{H}_0^1}\|\mathbf{v}\|_{\mathbf{H}_0^1}, \ \forall\, \mathbf{w}, \mathbf{v} \in \mathbf{H}_0^1. \tag{3.65}$$

Then, by the Babŭska-Lax-Milgram Theorem there exist a unique solution $\mathbf{w} \in \mathbf{H}_0^1(\Omega)$ of the system (3.55). Moreover, $\mathbf{w}$ depends continuously on the descent direction. □

From the previous Lemma, we can establish the following useful property of the second order term $\mathcal{G}$.

Corollary 3.2. *$\mathcal{G}$ satisfy the following pair of bounds:*

$$C\|\mathbf{w}\|_{\mathbf{H}_0^1}^2 \leq \langle \mathcal{G}\mathbf{w}, \mathbf{w}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} \leq \tilde{C}\|\mathbf{w}\|_{\mathbf{H}_0^1}^2. \tag{3.66}$$

Proof. The result follows from (3.60) and taking $\mathbf{v} = \mathbf{w}$ in (3.65). □

3.8 Exact Penalization Algorithm

Having discussed the main properties of the generalized derivatives of the objective functional, we introduce the following second–order algorithm for solving the *exact penalization* formulation (EP) numerically.

Algorithm 1: Exact Penalization Algorithm - Preliminar version

Initialize $\mathbf{u}_0$ such that $\operatorname{div}\mathbf{u}_0 = 0$ and set $k = 0$;
while *stopping criterion is not satisfied* **do**
  compute $\bar{\mathbf{d}}$ by solving problem (3.45);
  compute descent direction $\mathbf{w}_k$ by solving system (3.55);
  execute line–search to get α_k;
  update $\mathbf{u}_{k+1} := \mathbf{u}_k + \alpha_k \mathbf{w}_k$, and set $k = k + 1$.
end

From Algorithm 1 we can infer some useful properties of the approximated solution of problem (EP).

Lemma 3.6. *Let $\mathbf{w}_k$ be the descent direction at the k-th iteration of Algorithm 1, satisfying (3.55). Then, in the div-active set A_γ^k, $\mathbf{w}_k$ satisfies the following bound*

$$\|\operatorname{div}\mathbf{w}_k\|^2_{L^2(A_\gamma^k)} \leq \frac{1}{\gamma}\frac{\|\bar{\boldsymbol{d}}_k\|^2_{\mathbf{H}_0^1}}{C}.$$

Proof. In this proof we shall use portions of the proof of Lemma 3.5. By setting $\mathbf{w} = \mathbf{w}_k$ in equation (3.59), we obtain:

$$\langle \mathcal{H}\mathbf{w}_k, \mathbf{w}_k\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = \gamma\|\operatorname{div}\mathbf{w}_k\|^2_{L^2(A_\gamma^k)}. \tag{3.67}$$

By collecting (3.67) and (3.58), and using (3.55) with $\mathbf{w} = \mathbf{w}_k$, we obtain that:

$$\begin{aligned}\gamma\|\operatorname{div}\mathbf{w}_k\|^2_{L^2(A_\gamma^k)} &= \langle \mathcal{H}\mathbf{w}_k, \mathbf{w}_k\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}\\ &= -(\bar{\boldsymbol{d}}_k, \mathbf{w}_k)_{\mathbf{H}_0^1} - \langle \mathcal{J}\mathbf{w}_k, \mathbf{w}_k\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}\\ &\leq \|\bar{\boldsymbol{d}}_k\|_{\mathbf{H}_0^1}\|\mathbf{w}_k\|_{\mathbf{H}_0^1} - C\|\mathbf{w}_k\|^2_{\mathbf{H}_0^1}.\end{aligned}$$

In addition, from (3.56) we deduce that:

$$\begin{aligned}\gamma\|\operatorname{div}\mathbf{w}_k\|^2_{L^2(A_\gamma^k)} &\leq \frac{1}{C}\|\bar{\boldsymbol{d}}_k\|^2_{\mathbf{H}_0^1} - C\|\mathbf{w}_k\|^2_{\mathbf{H}_0^1}\\ &\leq \frac{1}{C}\|\bar{\boldsymbol{d}}_k\|^2_{\mathbf{H}_0^1}.\end{aligned}$$

Dividing by γ both sides, we get the desired result:

$$0 \leq \|\operatorname{div}\mathbf{w}_k\|^2_{L^2(A_\gamma^k)} \leq \frac{1}{\gamma}\frac{\|\bar{\boldsymbol{d}}_k\|^2_{\mathbf{H}_0^1}}{C}.$$

□

3.8.1 Discussion on the set $\Omega \setminus A_\gamma^k$

From Lemma (3.6), we established that the L^2-norm of the descent direction remains bounded over the active set A_γ^k by $\mathcal{O}(\gamma^{-1})$. Next we analyze the complement set, $\Omega\setminus A_\gamma^k$.

If $k = 0$, $\mathbf{u}_0$ is choosen such that $\operatorname{div}\mathbf{u}_0 = 0$, then

$$A_\gamma^0 = \left\{x \in \Omega : |\operatorname{div}\mathbf{u}_0(x)| < \frac{\sigma}{\gamma}\right\} = \Omega.$$

Therefore, by applying Lemma 3.6, we obtain the bound:

$$\|\operatorname{div} \mathbf{w}_0\|^2_{L^2(\Omega)} = \|\operatorname{div} \mathbf{w}_0\|^2_{L^2(A^0_\gamma)} \leq \frac{1}{\gamma} \cdot \frac{\|\bar{\boldsymbol{d}}_0\|^2_{\mathbf{H}^1_0}}{C}. \tag{3.68}$$

At $k = 1$ we have that

$$\operatorname{div} \mathbf{u}_1 = \operatorname{div} \mathbf{u}_0 + \alpha_0 \operatorname{div} \mathbf{w}_0,$$

since $\operatorname{div} \mathbf{u}_0 = 0$, we get that $\|\operatorname{div} \mathbf{u}_1\|_{L^2(\Omega)} = \alpha_0 \|\operatorname{div} \mathbf{w}_0\|_{L^2(\Omega)}$. Therefore, from (3.68) we obtain:

$$\|\operatorname{div} \mathbf{u}_1\|^2_{L^2(\Omega)} \leq \frac{\alpha_0^2}{\gamma} \cdot \frac{\|\bar{\boldsymbol{d}}_0\|^2_{\mathbf{H}^1_0}}{C}. \tag{3.69}$$

To analyze the descent direction $\mathbf{w}_1$, we recall from Lemma 3.6 that

$$\|\operatorname{div} \mathbf{w}_1\|^2_{L^2(A^1_\gamma)} \leq \frac{1}{\gamma} \cdot \frac{\|\bar{\boldsymbol{d}}_1\|^2_{\mathbf{H}^1_0}}{C}, \tag{3.70}$$

which shows that the L^2-norm of the divergence is also bounded on the active set A^1_γ. To address its behavior on the complement, we analyze the measure of $\Omega \setminus A^1_\gamma$. In this set we have that

$$\frac{\sigma}{\gamma} \leq |\operatorname{div} \mathbf{u}_1(x)|.$$

Squaring and integrating both sides over $\Omega \setminus A^1_\gamma$ yields:

$$\frac{\sigma^2}{\gamma^2} |\Omega \setminus A^1_\gamma| \leq \int_{\Omega \setminus A^1_\gamma} |\operatorname{div} \mathbf{u}_1(x)|^2 \, dx = \|\operatorname{div} \mathbf{u}_1\|^2_{L^2(\Omega \setminus A^1_\gamma)}. \tag{3.71}$$

Moreover, from (3.69) we deduce that (3.71) satisfies:

$$\frac{\sigma^2}{\gamma^2} |\Omega \setminus A^1_\gamma| \leq \|\operatorname{div} \mathbf{u}_1\|^2_{L^2(\Omega \setminus A^1_\gamma)} \leq \|\operatorname{div} \mathbf{u}_1\|^2_{L^2(\Omega)} \leq \frac{\alpha_0^2}{\gamma} \cdot \frac{\|\bar{\boldsymbol{d}}_0\|^2_{\mathbf{H}^1_0}}{C}, \tag{3.72}$$

which yields that

$$|\Omega \setminus A^1_\gamma| \leq \frac{\gamma}{\sigma^2} \cdot \frac{\alpha_0^2 \|\bar{\boldsymbol{d}}_0\|^2_{\mathbf{H}^1_0}}{C}, \tag{3.73}$$

this relation suggest choosig $\sigma^2 > \max\{\sigma_0^2, \gamma\}$ to control the size of $|\Omega \setminus A^1_\gamma|$. This becomes an error of approximation when discretized by the finite elements and eventually, the set $\Omega \setminus A^1_\gamma$ becomes empty if the mesh-size is smaller than the bound. This is observed in the numerical experiments along all iterates in Section 3.4

This discussion suggests using only the regular part of the objective functional to solve the discrete problem since the set $\Omega \setminus A^k_\gamma$ might become empty in the discrete domain if γ and σ are sufficiently large.

Henceforth, by the Riesz representation theorem, system (3.55) becomes:

$$\langle \mathcal{G}\mathbf{w}_k, \mathbf{v}_k \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = \langle -J'(\mathbf{u}_k), \mathbf{v}_k \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}, \ \forall \mathbf{v}_k \in \mathbf{H}_0^1(\Omega) \tag{3.74}$$

$$= -(\nabla J(\mathbf{u}_k), \mathbf{v}_k)_{\mathbf{H}_0^1}, \ \forall \mathbf{v}_k \in \mathbf{H}_0^1(\Omega) \tag{3.75}$$

Thanks to the previous discussion we propose the following modification to the previous version of the *exact penalization* Algorithm:

Algorithm 2: Exact Penalization Algorithm

Initialize $\mathbf{u}_0$ such that $\operatorname{div} \mathbf{u}_0 = 0$ and set $k = 0$;
while *stopping criterion is not satisfied* **do**
  compute the derivative of the regular part $-J'(\mathbf{u}_k)$ given by (3.11);
  compute the descent direction $\mathbf{w}_k$ by solving the system:
  $\langle \mathcal{G}\mathbf{w}_k, \mathbf{v}_k \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = \langle -J'(\mathbf{u}_k), \mathbf{v}_k \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1}, \ \forall \mathbf{v}_k \in \mathbf{H}_0^1(\Omega)$;
  execute line-search to get α_k;
  update $\mathbf{u}_{k+1} := \mathbf{u}_k + \alpha_k \mathbf{w}_k$, and set $k = k + 1$.
end

3.8.2 Discussion on the convergence of Algorithm 2

We do not provide a convergence result for Algorithm 2. This analysis is beyond the present work and will be addressed in future research. However, we discuss a possible approach to establish a convergence result by interpreting it as an inexact Semismooth Newton Method.

Recalling that the optimality condition associated with problem (EP) can be written as

$$\langle -J'(\bar{\mathbf{u}}_\sigma), \mathbf{v} \rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} = (\zeta, \operatorname{div} \mathbf{v})_{L^2} \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega), \tag{3.76}$$

where $\zeta \in \sigma \partial \| \cdot \|_{L^1}(\operatorname{div} \bar{\mathbf{u}}_\sigma)$. By the definition of the subdifferential of the L^1-norm, this condition implies:

$$\begin{cases} \zeta = \sigma \operatorname{sign}(\operatorname{div} \bar{\mathbf{u}}_\sigma) & \text{a.e. on } \{x \in \Omega : \operatorname{div} \bar{\mathbf{u}}_\sigma(x) \neq 0\}, \\ |\zeta| \leq \sigma & \text{a.e. on } \{x \in \Omega : \operatorname{div} \bar{\mathbf{u}}_\sigma(x) = 0\}. \end{cases} \tag{3.77}$$

Combining (3.76) and (3.77), we can reformulate the optimality system as a nonsmooth problem:

$$F(\mathbf{u}, \zeta) = \mathbf{0}, \tag{3.78}$$

where the operator $F : \mathbf{H}_0^1(\Omega) \times L^\infty(\Omega)$ is defined by

$$\begin{aligned} \langle -J'(\bar{\mathbf{u}}_\sigma), \mathbf{v}\rangle_{\mathbf{H}^{-1},\mathbf{H}_0^1} - (\zeta, \operatorname{div} \mathbf{v})_{L^2} = 0 \quad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega), \\ \operatorname{div} \bar{\mathbf{u}}_\sigma - \max\Big(0, \operatorname{div} \bar{\mathbf{u}}_\sigma + c(\zeta - \sigma)\Big) - \min\Big(0, \operatorname{div} \bar{\mathbf{u}}_\sigma + c(\zeta + \sigma)\Big) = 0 \quad \text{a.e. in } \Omega, \end{aligned} \tag{3.79}$$

for some fixed constant $c > 0$, see [106].

The semismoothness of F might be analyzed in infinite or in finite dimensions after discretization. Solving (3.79) in functional spaces requires a more in-depth analysis of the differentiability of max and min functions, considering the differentiability gap in $L^p(\Omega)$ spaces, similarly to Section 3.7.2. In addition, the generalized Newton differentiability of all involved functions and the solvability of the resulting system have to be ensured. Then, a comparison with system (3.75) could be performed to match the equations and identify the sources of inexactness. In this way, the Algorithm 2 can be cast as an Inexact Newton method, for which convergence is known.

Thus, the convergence analysis of Algorithm 2 might be carried out in the framework of the Inexact Semismooth Newton method.

3.8.3 Line–search routine

The algorithm stops when the norm of the difference of two consecutive approximated solutions drops below a given tolerance, i.e., $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{H}_0^1}$ serves as a descent indicator and stopping criteria when it is approximately zero.

Additionally, the election of the step length α is key to guarantee the sufficient decrease. In the *Exact Penalization* Algorithm 2- step 5 we use a line search technique which exploits polynomial models of the objective function for backtracking. This stepsize reduction approach was proposed in [43, Sec. 6.3.2]. If a stepsize does not satisfy the sufficient decrease condition, the next candidate will be constrained in an interval that depends on the previous stepsize. Hence, we have:

$$\alpha_k \in [c_l \alpha_{prev}, c_u \alpha_{prev}], \quad \text{for} \quad k = 0, \cdots$$

where c_l and c_u are positive constants and α_{prev} stands for the previous step length value. In general, it is mandatory to construct stepsizes that are bounded away from zero [77, Sec. 3.2].

To finish this section, we state the following result with some convergence related quantities. This is a direct consequence of the argumentation in Section 3.8.1 and Corollary 3.2.

Corollary 3.3. *Let* $\mathbf{w}_k$ *satisfy* (3.75). *Then the following inequalities hold:*

(i) $C\|\mathbf{w}_k\|^2_{\mathbf{H}^1_0} \leq -\langle J'(\mathbf{u}_k), \mathbf{w}_k\rangle_{\mathbf{H}^{-1},\mathbf{H}^1_0} \leq \tilde{C}\|\mathbf{w}_k\|^2_{\mathbf{H}^1_0}$,

(ii) $\langle J'(\mathbf{u}_k), \mathbf{w}_k\rangle_{\mathbf{H}^{-1},\mathbf{H}^1_0} < 0$,

(iii) $C\|\mathbf{w}_k\|_{\mathbf{H}^1_0} \leq \|J'(\mathbf{u}_k)\|_{\mathbf{H}^{-1}}$.

Proof. This follows simply by taking $\mathbf{v}_k = \mathbf{w}_k$ in (3.75) and from Corollary 3.2. □

3.9 Numerical Experiments

The final section of this chapter is devoted to the numerical experimentation of Algorithm 2. In the first subsection, we conduct a set of experiments linked to the algorithm's performance, including an exhaustive testing of the parameters governing the associated regularizations and, in particular, to the exact penalization parameter σ. The second set of experiments, in Section 3.9.2, aims to compare our algorithm with the Semismooth Newton method, which is well known by its superlinear convergence properties. A third set of experiments in three dimensions are also presented to further illustrate the applicability and scalability of the exact penalization method which will be referred as EP algorithm.

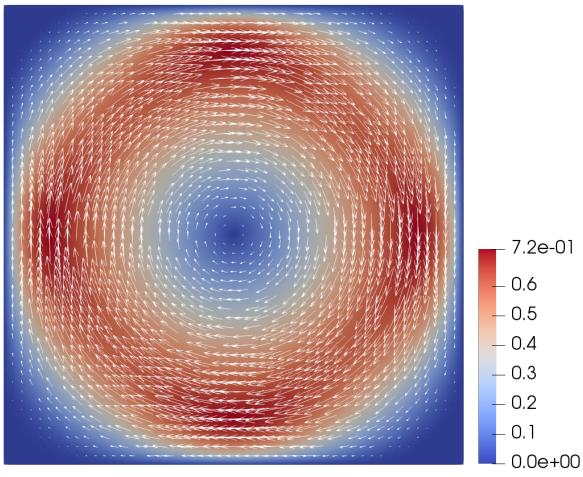
Implementation details

We consider an open subset Ω of $\mathbb{R}^n$, with $n = 2, 3$, which is a polygonal/polyhedral domain and $\mathcal{T}_h$ a regular discretization (by triangles or tetrahedrons) on Ω. The Galerkin Finite Element Method was used to approximate the desired velocity $\bar{\mathbf{u}}_\sigma \in \mathbf{H}^1_0$ - of the Bingham problem - by continuous piecewise quadratic vector-valued Lagrange trial functions over each conforming finite element. Therefore, we consider the finite-element space $V_h = \{\mathbf{u}_h \in C(\bar{\Omega})^2 |\ \mathbf{u}_h|_\Gamma = 0 \text{ and } \mathbf{u}_h|_T \in P^2, \forall T \in \mathcal{T}_h\}$. Our experiments were implemented in the open-source software FEniCS (https://fenicsproject.org/).

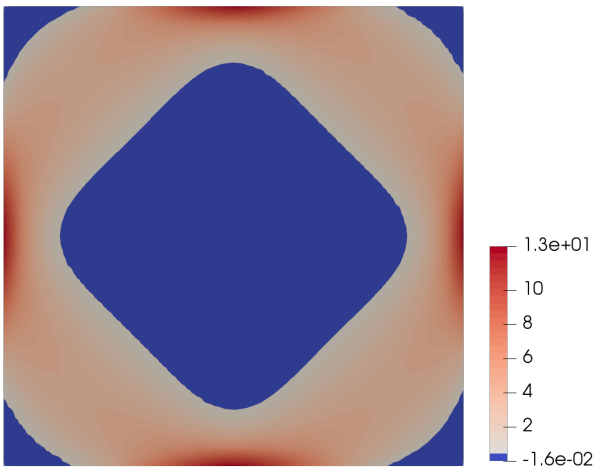
3.9.1 Algorithm's performance

In order to measure the performance of the EP algorithm, we consider the benchmark of a rotational Bingham flow in a square reservoir. In this case, problem (EP) is solved with a driven force $\mathbf{f_b}(x_1, x_2) = 300(x_2 - 0.5, 0.5 - x_1)$ over the open subset $\Omega = (0, 1)^2 \subset \mathbb{R}^2$ with $g = 10\sqrt{2}$. In the following experiments the mesh size is $h = 1/50$. Figure 1 shows the computed rotational flow expected from applying the force f, whereas in Figure 3.2b we can visualize the yielded (dark blue) and unyielded regions (red tones). This configuration presents a central solid region.

Recall that Algorithm 2 has three important parameters, namely:



(a) Velocity field $\mathbf{u}$



(b) Plug zones given by $|\mathcal{E}(\mathbf{u})|_{\mathbb{L}^2}$

Figure 3.2: Bingham flow in the square reservoir.

- σ: exact penalization parameter, satisfying $\sigma \geq \sigma_0$, see Theorem 3.1
- γ: enriching second–order information parameter
- β : Huber regularization parameter

Experiment 1: test varying σ, γ and β

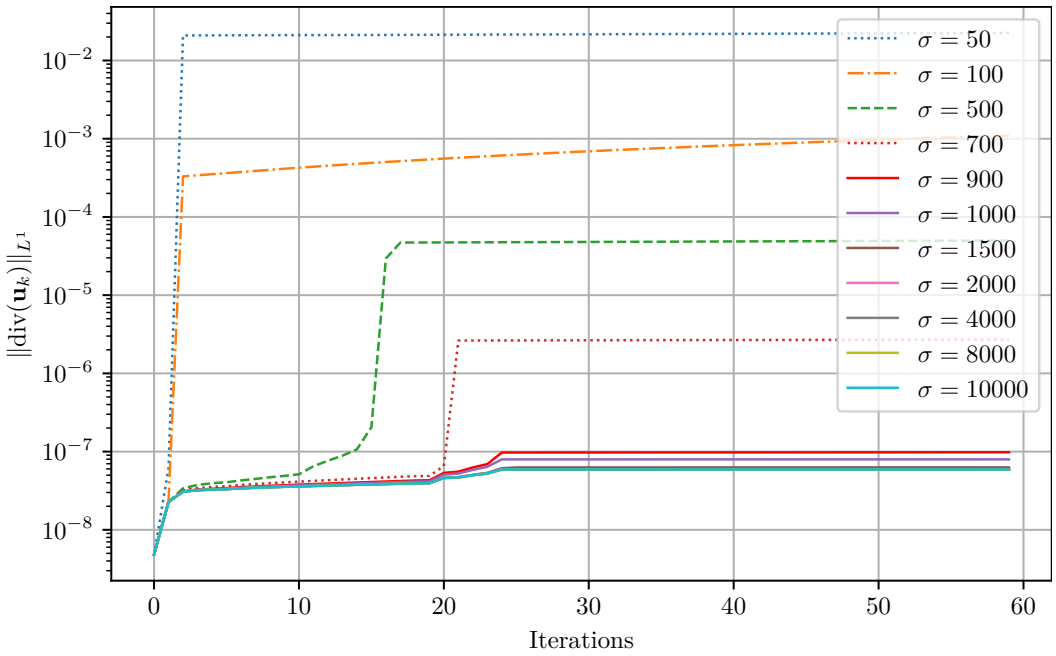
Table 3.1 sumarizes the influence of σ, γ and β on the numerical realization of the method. Here, the algorithm is terminated as soon as the stopping criteria is satisfied. That is, the quantity $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{H}_0^1}$ reaches a value below the tolerance $1e-5$. The computation time is also reported. *Parameter* σ : Figures 3.3a - 3.4a and Table 3.1 show the numerical behavior of choosing different values of σ for divergence term $\|\operatorname{div}\mathbf{u}\|_{L^1}$. In Figure 3.3a, the divergence is depicted for the first 60 iterations of the EP - Algorithm. We can observe that the history of the divergence values remain between $1.0e-7$ and $1.0e-8$ for values of $\sigma \geq 900$ (colored - solid lines), recognisably lower than those for smaller values of $\sigma < 900$ in dashed lines. This illustrates the equivalence of the exact penalization problem (EP) with the constrained formulation (CP). In previous sections we have discussed that this equivalence is given for all $\sigma \geq \sigma_0$, with $\sigma_0 \approx \|\lambda\|_{L^2}|\Omega|^{\frac{1}{2}}$. Despite the computation of $\|\lambda\|_{L^2}$ can not be a–priorily done, i.e. without an approximate solution at hand, we can confirm numerically that our estimation of σ_0 in Remark 3.3 is sharp. Indeed, we have estimated the upper bound: $896.5 \geq \|\lambda\|_{L^2}|\Omega|^{-1/2} \approx \sigma_0$. In Table 3.1, for variation of $\sigma \geq \sigma_0$, we observe that once σ is suitably chosen, its variation does not have a hard influence in the numerical performance of the EP-algorithm. Figure 3.4a shows this behavior for the cost funcional.

Parameter γ: from our theory, we know that in order to get a good approximation for the second–order information, γ must be sufficiently large, see Corolary **??**. As shown in Table 3.1, this parameter is crucial for our algorithm; in fact, if we neglect the additional second–order information ($\gamma = 0$) the algorithm fails to converge and we only display the results for the iteration $k = 100$. In contrast, by setting increasing values of $\gamma = 1e+8$ and $1e+9$ we see that $\|\operatorname{div}\mathbf{u}_k\|_{L^1}$ gets smaller. Notice that $\gamma = 1e+9$ achieves, on average, a divergence norm of order $6e-8$ and the algorithm seems to have a more stable values of the cost-functional $J(\mathbf{u}_k)$.

Also, we observe in the fifth column that the stopping criteria, $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{H}_0^1}$, (subsection 3.8.3) gets closer to zero for $\gamma = 1e+9$ (see Figure 3.4b). Based on these numerical results, we conclude that the additional second–order information, enriching the descent direction system (3.55), is essential for the method.

Numerical performance for the EP-Algorithm							
γ	β	σ	k	$\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{H}_0^1}$	$\|\operatorname{div} \mathbf{u}_k\|_{L^1}$	$J(\mathbf{u}_k)$	time (s)
0	500	900	100	0.048	7.55e-05	0.06	91.2
		2000	100	0.048	7.55e-05	0.14	91.2
		4000	100	0.048	7.55e-05	0.29	91.1
		8000	100	0.048	7.55e-05	0.59	91.2
		10000	100	0.048	7.55e-05	0.74	91.2
	1000	900	100	0.024	3.78e-05	0.03	91.4
		2000	100	0.024	3.78e-05	0.07	91.2
		4000	100	0.024	3.78e-05	0.14	91.2
		8000	100	0.024	3.78e-05	0.29	91.2
		10000	100	0.024	3.78e-05	0.37	91.0
1e+8	500	900	12	1.17e-06	7.58e-07	-8.41	7.8
		2000	12	1.14e-06	5.65e-07	-8.41	8.3
		4000	12	1.16e-06	5.64e-07	-8.40	8.2
		8000	12	2.45e-06	5.22e-07	-8.39	8.6
		10000	11	2.60e-06	5.11e-07	-8.39	7.9
	1000	900	28	1.33e-06	7.75e-07	-8.33	17.5
		2000	28	3.62e-07	5.90e-07	-8.32	18.3
		4000	37	8.05e-07	5.54e-07	-8.32	24.3
		8000	38	8.77e-07	5.40e-07	-8.31	25.2
		10000	27	2.22e-06	5.19e-07	-8.30	18.1
1e+9	500	900	14	8.87e-07	1.05e-07	-8.42	9.0
		2000	15	3.44e-07	7.03e-08	-8.42	9.9
		4000	15	6.66e-07	6.60e-08	-8.42	11.2
		8000	14	8.67e-07	6.33e-08	-8.42	9.7
		10000	14	8.63e-07	6.33e-08	-8.42	9.9
	1000	900	25	3.32e-07	9.73e-08	-8.33	22.6
		2000	26	3.11e-07	6.21e-08	-8.33	24.3
		4000	25	3.28e-07	5.91e-08	-8.33	23.6
		8000	25	3.31e-07	5.91e-08	-8.33	24.0
		10000	25	3.29e-07	5.91e-08	-8.33	24.1

Table 3.1: Performance of the EP- Algorithm for different values of γ, β and σ



(a) Divergence history, $\gamma = 1e+9$, $\beta = 1e+3$. Solid lines for $\sigma \geq 900$

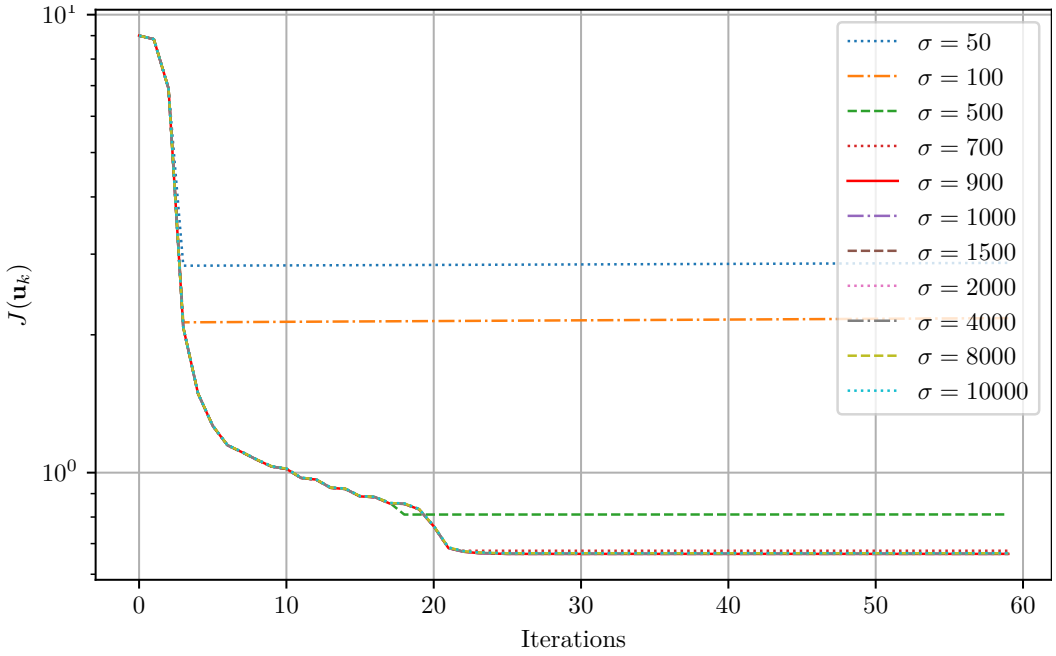
Figure 3.3: Experiment 1: Velocity's divergence in L^1-norm

3.9.2 Comparison with Newton Semismooth Method

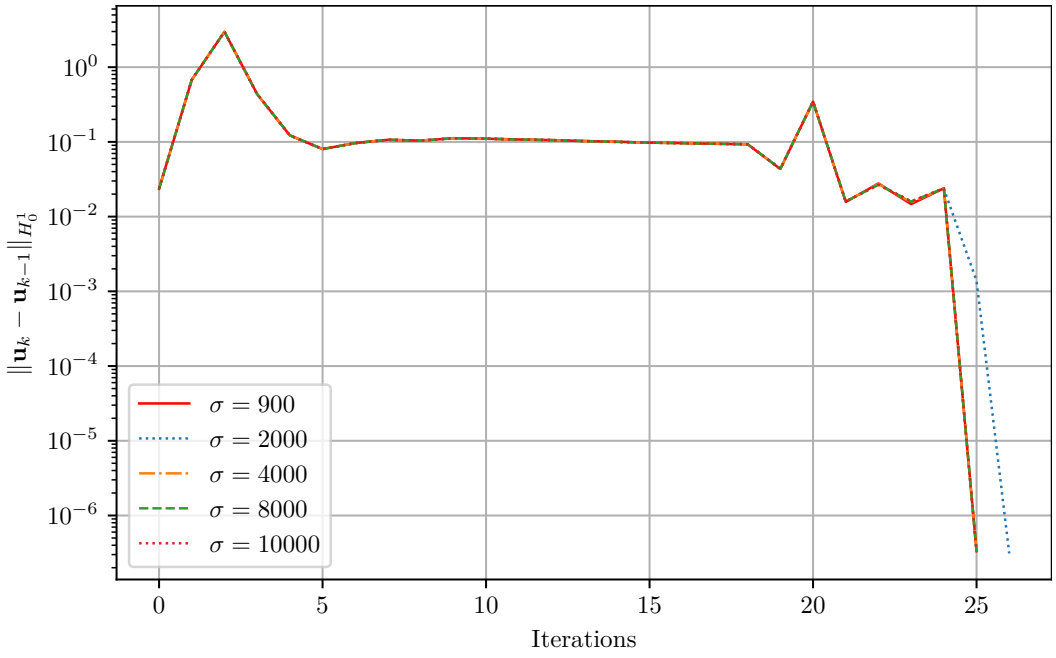
We compare the exact penalization method with the Newton semismooth method (SSN), which is also build on second-order information basis, and it is very well known by its superlinear rate of convergence properties (see [115, Ch.3]). Both methods are applied to the same triangulation of conforming piecewise quadratic finite elements. Notice that SSN solves the same regularized problem in a context of a nonlinear system (see [38]). Since the stopping criteria for each strategy differs, we compare a fix number of iterations of each algorithm. In the following set of experiments we illustrate that the exact penalization method can deliver several numerical benefits and drawbacks comparing with SSN. For further comparison, we also consider the quadratic penalization (QP).

Experiment 2: convergence to a Bingham's analytical solution

The original problem was regularized using the Huber–smoothing, depending on the parameter β. In the following experiment, the analytical solution of problem (3.5) of a fluid flow between two parallel plates is known, which allow us to compute the *error*= $\|\mathbf{u}_{exact} - \mathbf{u}_k\|_{\mathbf{H}^1_0}$ at each k-th iteration for EP and SSN algorithms. Here, the velocity field $\mathbf{u} = (\mathbf{u}_1(y), \mathbf{0}, \mathbf{0})$ is a scalar field that depends only on y in the x-direction. The corresponding minimization functional is simplified since the strain-rate tensor is given by the gradient ∇, i.e., we have to minimize $J(\mathbf{u}) := \frac{\mu}{2}\int_\Omega (\nabla\mathbf{u}, \nabla\mathbf{u})\, dx + g\int_\Omega |\nabla\mathbf{u}|\, dx -$



(a) Cost functional for several σ values



(b) Stopping criteria for several σ values

Figure 3.4: Experiment 1: EP performance with $\gamma = 1e + 9$ and $\beta = 1e + 03$

$\int_\Omega \mathbf{f_b} \cdot \mathbf{u}\, dx$. The analytical solution is given by $\mathbf{u}_{exact} = (\boldsymbol{u}_1(y), 0, 0)$ (see [2, Sec. 6.1.1.1]), where:

$$\boldsymbol{u}_1(y) = \begin{cases} \frac{1}{8}[(1-2g)^2 - (1-2g-2y)^2], & \text{if } 0 \le y < \frac{1}{2} - g, \\ \frac{1}{8}(1-2g)^2, & \text{if } \frac{1}{2} - g \le y \le \frac{1}{2} + g \\ \frac{1}{8}[(1-2g)^2 - (2y-2g-1)^2], & \text{if } \frac{1}{2} + g < y \le 1, \end{cases}$$

and the pressure drop $\mathbf{f_b}$, is given by $\mathbf{f_b} = -x$. Here, we set $g = 0.3$.

Following Remark 3.3, we chose $\sigma_0 \approx \|\lambda\|_{L^2}|\Omega|^{1/2} \le 17$ to guarantee the equivalence with the exact penalization formulation. For fixed $\gamma = 1e+09$, we solve the problem varying β as shown in Table 3.2. The Huber regularization of the Bingham term $g\int_\Omega |\mathcal{E}\mathbf{u}(x)|\, dx$ does not demand large values of β. Numerical experimentation of SSN method in [38] has shown that $\beta = 1e{+}03$ is large enough to get a local regularization of the nondifferentiable term. Therefore, in Table 3.2 we see no significant improvements in the *error* and cost functional for values of $\beta \ge 1e+03$ in the SSN method.

However, for the EP, as β increases the *error* decays accordingly and is lower than the SSN.

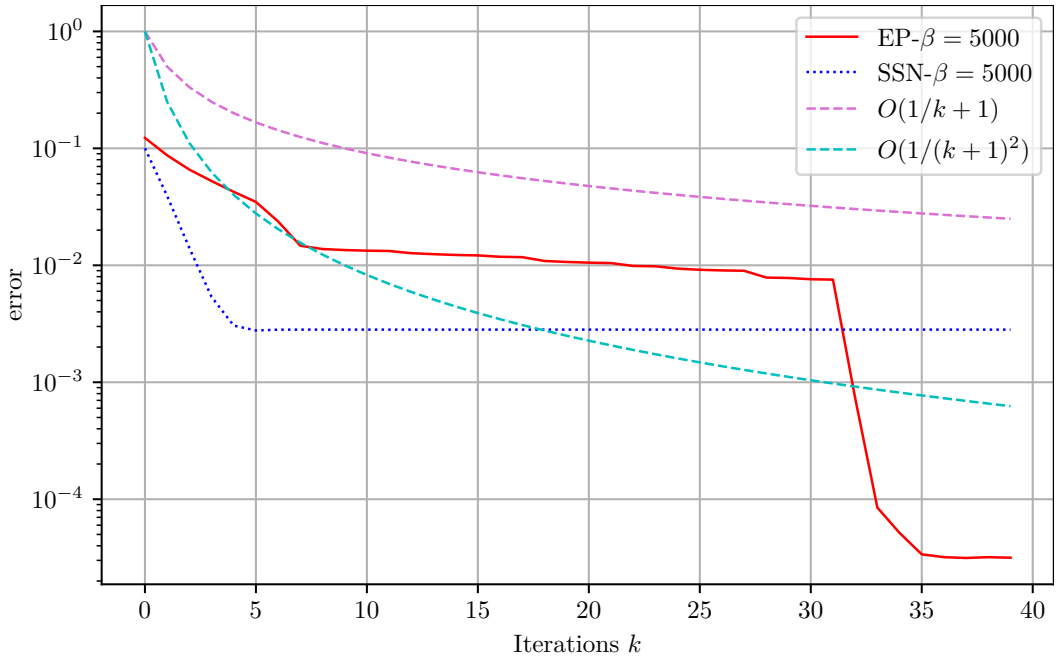
In Figure 3.5a, at first sight we observe that SSN is faster in the first iterations. However, we realize that the exact penalization second-order method continues to decrease the *error* with a pronounced fall in the last iterations, achieving a considerably lower error and cost (see Table 3.2 - seventh column) compared with SSN. Furthermore, the exact penalization algorithm computes approximate solutions with a more precise divergence.

Taking into account the relation $\|\operatorname{div}\mathbf{u}_k\|_{L^1} \le |\Omega|^{1/2}\|\operatorname{div}\mathbf{u}_k\|_{L^2}$, we chose the L^2–norm of the divergence as a reference for all methods. We observe that the L^2–norm of the divergence is effectively smaller for EP method. Here, we see the advantage of the L^1–norm penalization of the divergence term against the SSN method, where the equation involving the divergence of the velocity is formulated in the space L^2 which do not induce sparsity.

On the other hand, we confirm numerically (in Table 3.2 columns fifth - sixth) that the L^1–norm promotes sparsification of the divergence term. Moreover, in the SSN method, the variation of parameter β does not lead to a decrease in the divergence norm because the underlying system is decoupled and the equation involving divergence of the velocity and the pressure p is stabilized by the small parameter $\varsigma > 0$ (see [38, Sec 6.1]), i.e., the incompressibility constraint is relaxed by $\varsigma = 1e-05$. Therefore, parameter β has no influence over this equation.

β		$\|\mathbf{u}_{exact} - \mathbf{u}_k\|_{\mathbf{H}^1_0}$		$\|\operatorname{div} \mathbf{u}_k\|_{L^2}$		$J(\mathbf{u}_k)$		time (s)	
	k	EP	SSN	EP	SSN	EP	SSN	EP	SSN
100	1	0.121	0.100	1.28e-13	2.46e-08	0.032	4.79e-03	1.19	1.46
	10	0.015	0.002	9.59e-12	3.21e-08	-2.43e-03	-2.75e-03	14.88	15.62
	20	1.34e-03	2.99e-03	9.92e-12	3.21e-08	-2.93e-03	-2.75e-03	26.80	31.34
	30	1.34e-03	2.99e-03	9.92e-12	3.21e-08	-2.93e-03	-2.75e-03	39.60	46.35
	40	1.34e-03	2.99e-03	9.92e-12	3.21e-08	-2.93e-03	-2.75e-03	52.91	64.63
500	1	0.121	0.100	2.05e-13	2.49e-08	0.033	0.005	1.16	1.33
	10	0.013	2.82e-03	4.94e-11	6.35e-08	-2.40e-03	-2.42e-03	15.32	14.477
	20	2.2e-03	2.82e-03	5.44e-11	6.35e-08	-2.70e-03	-2.42e-03	27.83	28.614
	30	2.68e-04	2.82e-03	5.42e-11	6.35e-08	-2.70e-03	-2.42e-03	40.24	42.795
	40	2.67e-04	2.82e-03	5.42e-11	6.35e-08	-2.70e-03	-2.42e-03	52.61	57.00
1000	1	0.121	0.100	2.53e-13	2.49e-08	0.033	0.005	1.16	1.29
	10	0.013	2.81e-03	3.44e-11	6.48e-08	-2.29e-03	-2.38e-03	14.94	14.45
	20	7.41e-03	2.81e-03	3.53e-11	6.48e-08	-2.65e-03	-2.38e-03	27.63	28.71
	30	1.34e-04	2.81e-03	3.54e-11	6.48e-08	-2.69e-03	-2.38e-03	40.28	42.99
	40	1.34e-04	2.81e-03	3.54e-11	6.48e-08	-2.69e-03	-2.38e-03	52.69	57.31
5000	1	0.122	0.100	5.27e-13	2.49e-09	0.033	0.005	1.14	1.30
	10	0.013	2.81e-03	4.32e-11	6.48e-08	-2.39e-02	-2.34e-03	15.19	14.89
	20	0.010	2.81e-03	4.57e-11	6.48e-08	-2.59e-03	-2.34e-03	28.29	29.43
	30	7.79e-03	2.81e-03	4.67e-11	6.48e-08	-2.63e-03	-2.34e-03	42.35	43.77
	40	3.16e-05	2.81e-03	5.36e-11	6.48e-08	-2.67e-03	-2.34e-03	56.34	58.110

Table 3.2: Experiment 2: first 40 iterations of EP vs SSN.



(a) Error decay comparison between EP and SSN

Figure 3.5: Experiment 2 - flow between two plates

Experiment 3: Comparison with Second-order Methods

In this experiment we compare three methods based on second other information: SSN, EP and Quadratic penalization (QP) for the benchmark rotational flow test presented in section 3.9.1. We compare the first 100 iterations of each algorithm in Table 3.3. The EP-algorithm is competitive with SSN and the Quadratic penalization method since no major oscillations in the algoritms's performance are shown after the 25th iteration for the three strategies. However, the advantage of the exact penalization is evident when reaching the restriction $\|\operatorname{div} \mathbf{u}_k\|$ close to zero. Observe in Figure 3.6b that, for several σ values, the lowest magnitud for the velocity divergence is achieved by the EP-algorithm. This is somehow expected by sparsification properties of the L^1–norm. For the same σ values the quadratic method hardly achieves an order of $1e-04$. And, in the case of SSN, the velocity divergence norm is greater than in the EP-Algorithm. This behavior is also depicted in Table 3.3 for the L^2–norm of the velocity divergence of the approximated solution.

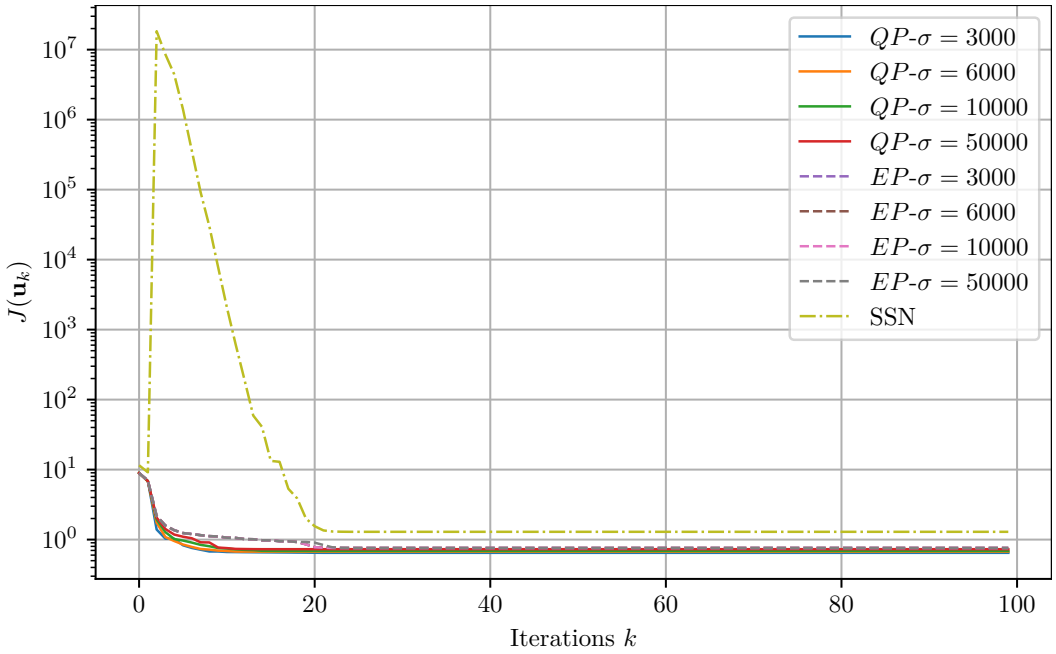
The decay of the objective function is plotted in Figure 3.6a. Here, EP-algorithm and QP-algorithm attain smaller values than the SSN method.

k	$\|\operatorname{div} \mathbf{u}_k\|_{L^2}$			$J(\mathbf{u}_k)$			time (s)		
	EP	QP	SSN	EP	QP	SSN	EP	QP	SSN
1	7.170e-09	0.00167	0.000413	-0.159	-0.170	2.406	0.59	0.652	0.7827
10	4.576e-08	0.00277	0.0140	-7.890	-8.417	7811.114	5.59	5.34	14.75
20	5.056e-08	0.00274	1.122e-04	-8.236	-8.443	-7.094	11.96	11.35	22.65
25	7.485e-08	0.00273	7.836e-05	-8.447	-7.807	-7.708	15.08	14.25	26.42
30	7.486e-08	0.00273	7.808e-05	-8.335	-8.447	-7.808	18.54	17.18	30.14
50	7.487e-08	0.00273	7.808e-05	-8.335	-8.447	-7.808	32.53	28.89	45.22
80	7.490e-08	0.00273	7.808e-05	-8.335	-8.447	-7.808	53.72	46.59	67.88
100	7.491e-08	0.00273	7.808e-05	-8.335	-8.447	-7.808	67.93	58.76	83.15

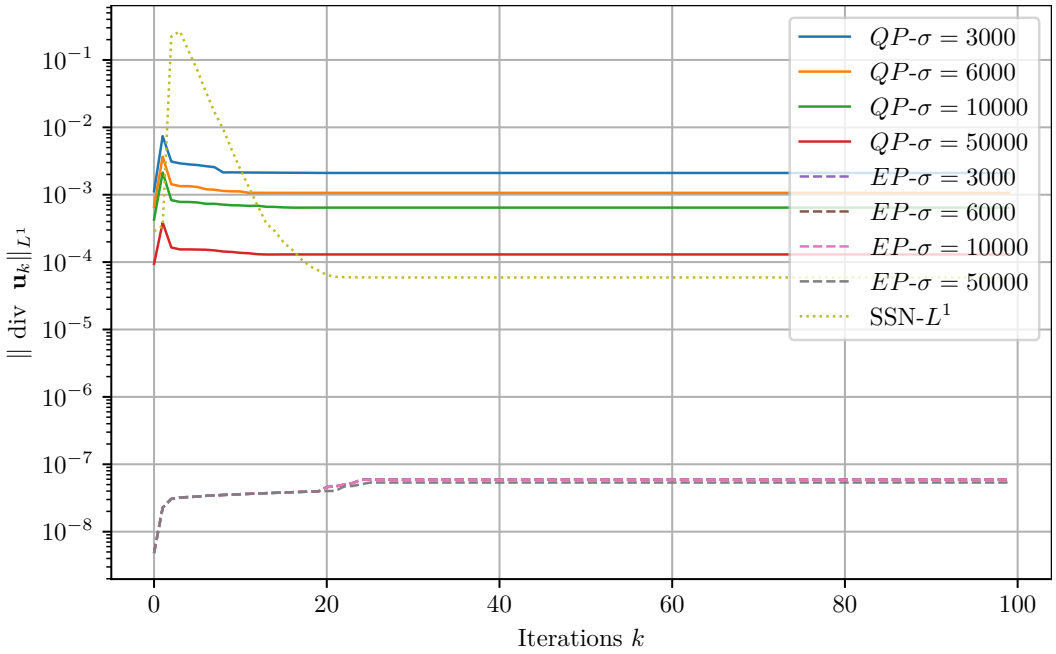
Table 3.3: Experiment 3: comparison between EP, QP and SSN algorithms with parameters: $\gamma = 1e+9$, $\beta = 1e+3$ and $\sigma = 3e+3$

3.9.3 Numerical Experiments in 3D Geometries

In this section we examine the EP-algorithm in 3D geometries. We take advantage of FEniCS versatility for testing the EP-algorithm using three-dimensional finite elements described in the Implementation Details, and the FEniCS parallelization capabilities for solving the associated variational problems within the algorithm. We run these tests on a high performance computing system HP ProLiant BL460c Gen8.



(a) Cost in logarithmic scale



(b) Comparison of the divergence history of approximated solutions in L^1-norm

Figure 3.6: Experiment 3 - comparison between EP, QP and SSN algorithms

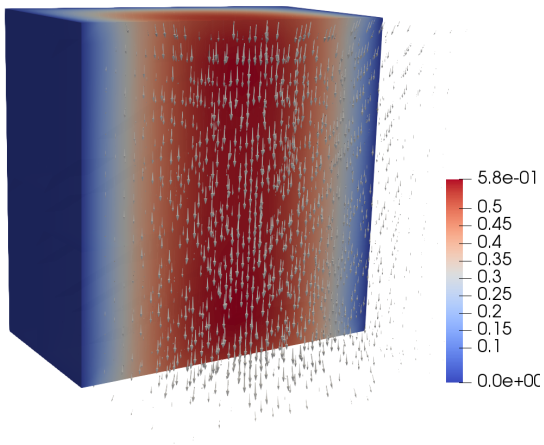


Figure 3.7: Velocity field

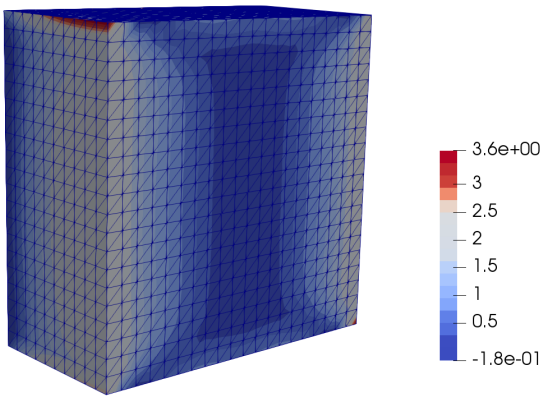


Figure 3.8: Plug flow

Experiment 4: Cube

In this experiment we consider a Bingham fluid in a cubic geometry. We assume a laminar flow with constant drop pressure c along the z-axis. The drop in pressure is considered in the periodic boundary conditions of the model (see [81, Sec. 6.2]). Therefore, the associated minimization problem reads as:

$$J = \mu \int_\Omega \mathcal{E}\mathbf{u} : \mathcal{E}\mathbf{u}\, dx + \int_\Omega \Psi(\mathcal{E}\mathbf{u})\, dx - \int_\Omega \mathbf{f_b}\mathbf{u}\, dx - \int_\Gamma c\mathbf{n}|_{z=0}\mathbf{u}|_{z=0}\, ds + \sigma \| \operatorname{div} \mathbf{u}\|_{L^1}$$

here $\Gamma = [0,1] \times [0,1] \times \{0\}$, $\mathbf{f_b} = 0$, $c = 10$, $g = 0.5\sqrt{2}$, $\beta = 1e+3$, $\gamma = 1e+7$ and $\sigma = 1e+4$. The unit cube is discretized with tetrahedrons with step size $h = 1/20$. Figure 3.7 shows the velocity field. In the center of the cube the fluid acts like a rigid material as well as the corners of the cube. In Figure 3.8 the Frobenius norm $|\mathcal{E}\mathbf{u}|$ and the discretization are depicted in the geometry cut by a parallel plane to the y axis. Here, the plug zones are colored in light blue. The algorithm performed 4 iterations with the following values $J_\sigma = -1.098$, $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{H}_0^1} = 5.614e-07$, and $\| \operatorname{div} \mathbf{u}\|_{L^1} = 1.57e-06$. Taking advantage that FEniCS run in parallel using MPI and without modifying the algorithm, in Table 3.4 we report on the comparison between parallel runtime in several CPUs for EP and SSN algorithms. Despite the efficiency of the current implementation deteriorates, the time reduction is significant when more CPUs are added. Also, its execution time escalates better that the execution time in SSN. However, the percentage of time reduction is similar for both methods as more CPUs are incorporated to the computation process. Because of memory limitations, there was not possible to run the experiment with the mesh size $h = 1/20$ in one core. Therefore, we calculate the speedup and the efficiency with the execution time reference in 2 cores.

No. cores	Time (s)		% time reduction		Speedup		Efficiency	
	EP	SSN	EP	SSN	EP	SSN	EP	SSN
2	1218.2	5359.87	100%	100%	1	1	1	1
6	709.8	2828.89	58.2%	52.7%	1.71	1.89	0.28	0.31
12	395.8	2123.41	32.4 %	39.6%	3.07	2.52	0.25	0.21
24	240.9	1619.87	19.7%	30.2%	5.05	3.30	0.21	0.13

Table 3.4: Experiment 4: EP vs. SSN algorithms scaling performance

Experiment 5: Lid-driven cavity

Now, we test a lid-driven viscoplactic flow inside a unite cube. The geometry is discretized with tetrahedrons with step size $h = 1/30$. The corresponding body force is given by $\mathbf{f_b}(\mathbf{x}) = \mathbf{0}$, since the motion is given by a moving lid, i.e., we have $\mathbf{u}_D(\mathbf{x}) = (1,0,0)^\top$ if $x_3 = 1$ and $\mathbf{u}_D(\mathbf{x}) = (0,0,0)^\top$, otherwise. These boundary conditions add a new constraint to our optimization problem, i.e. $\min_{\mathbf{u}\in\mathbf{H}^1(\Omega)} J(\mathbf{u})$ subject to $\mathbf{u} = \mathbf{u}_D$ on $\partial\Omega$.

To cope with this new constraint, let us fix $\mathbf{u_0} \in U = \{\mathbf{u} \in \mathbf{H}^1(\Omega) | \operatorname{div}(\mathbf{u}) = 0, \mathbf{u}|_{\partial\Omega} = \mathbf{u}_D\}$. The solution $\bar{\mathbf{u}}$ is given by $\bar{\mathbf{u}} = \mathbf{u_0} + \hat{\mathbf{u}}$, where $\hat{\mathbf{u}}$ is the minimizer of

$$\mu \int_\Omega \mathcal{E}(\mathbf{u}+\mathbf{u_0}) : \mathcal{E}(\mathbf{u}+\mathbf{u_0})\, dx + \int_\Omega \Psi(\mathcal{E}(\mathbf{u}+\mathbf{u_0}))\, dx - \int_\Omega \mathbf{f_b}(\mathbf{u}+\mathbf{u_0})\, dx + \sigma \| \operatorname{div}(\mathbf{u})\|_{L^1}.$$

The parameters have the following setting: $g = 2$, $\beta = 1e+3$, $\gamma = 1e+9$, $\mu = 0.5$ and $\sigma = 1e+4$. Figure 3.9 show the velocity field of the fluid in the cube cut by half in the y axis. The fluid rotates in the interior of the geometry however, thanks to this rotation the material moves without continuous deformation in the center of the cube. Figure 3.10 shows the plug zones in light blue. The numerical performance of the algorithm is displayed in Table 3.5. The divergence norm, the number of iterations and the stopping criteria behave similar to the 2D case. For instance $\| \operatorname{div} \mathbf{u}_k\|_{L^1}$ achieves an order of $e-08$ and the stopping criteria $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{H}^1_0}$ drops close to $e-07$.

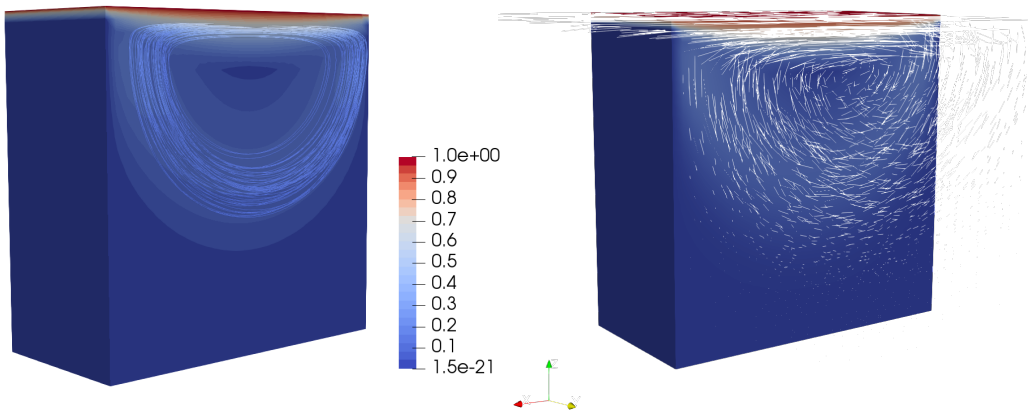


Figure 3.9: Experiment 5: stream lines and velocity field of lid-driven flow

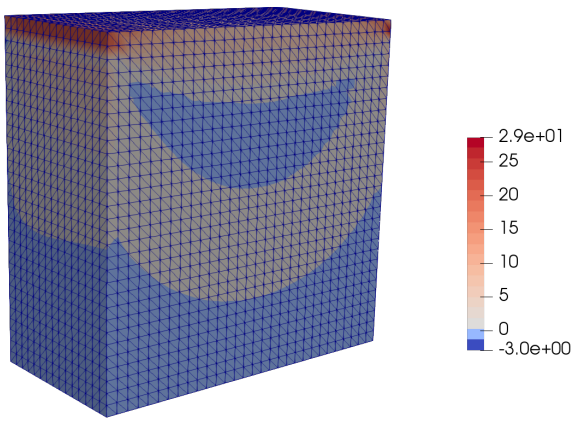


Figure 3.10: Experiment 5: plug zones

Exact Penalization			
k	$\|\operatorname{div} \mathbf{u}_k\|_{L^2}$	$J(\mathbf{u}_k)$	$\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_{\mathbf{H}_0^1}$
1	2.58e-09	5.39	0.476
3	6.07e-09	5.17	0.254
5	9.50e-09	5.10	0.087
8	1.94e-08	5.08	0.030
10	2.22e-08	5.08	6.47e-03
13	2.41e-08	5.08	1.04e-03
15	2.52e-08	5.08	3.76e-05
18	2.52e-08	5.08	3.60e-07
20	2.52e-08	5.08	3.60e-07

Table 3.5: Experimemt 5: 3D Lid-driven cavity with $g = 2, \beta = 1e + 3, \gamma = 1e + 9$

Chapter 4

Part II: Group-Sparse Optimization Methods with Applications to Bingham Fluids and Non-Convex Optimization

In this chapter, we developed optimization methods tailored for the numerical solution of group-sparse problems. These problems are characterized by the incorporation of the $\|\cdot\|_{1,2}$ norm, which is defined as the sum of Euclidean norms. As a regularizer, this norm promotes sparsity across groups of variables rather than individual variables.

We build on the strategies developed in Chapter 3, incorporating second-order information to accelerate the search direction. In addition, we propose an active-set strategy to *a priori* determine the sparse groups of variables. The novelty of the active-set phase lies in its dynamic and efficient identification of sparse groups through an iterative analysis of the angle of two consecutive iterations. In some applications, the active-set phase reduces the size of the Newton-type system by using the curvature information exclusively on the non-sparse or inactive groups.

This chapter is divided into two main sections. In the first one, Section 4.2, we develop a group-sparse algorithm specifically tailored for solving the Bingham flow problem in a pipe without any regularization procedure. This unregularized and simplified problem is equivalently reformulated as a linearly constrained minimization problem by means of a dual variable. The new constrained term, in terms of the dual variable, is interpreted as a group sparsity regularizer in the regions where the material behaves like a rigid solid. We discretized the problem using the Galerkin Finite Element Method and analyze its augmented Lagrangian formulation. The further subsections are devoted to the construction of the algorithm. First, we derived the steepest de-

scent direction, which is determined using the directional derivative. The directional derivative is computed separately for the primal and dual variables. Furthermore, due to the group structure of the dual variable, the steepest descent direction is calculated in a group-wise manner. In Section 4.2.3 we incorporate second-order information and derive a Newton-type system. Section 4.2.5 is devoted to the analysis of the active-set strategy, which is determined based on the angle between two consecutive iterations. If this angle is obtuse, the iterate is projected to zero. Otherwise, the iterate is updated by the steepest descent direction. We conclude the first section with numerical experiments to evaluate the performance of the algorithm.

In the second section, 4.3, building on the concepts introduced for the Bingham group-sparse algorithm of the first part, we extend these strategies to address a more general group-sparse optimization problem of the form:

$$\min_{u\in\mathbb{R}^m} f(\mathbf{u}) + \sigma\|\mathbf{u}\|_{1,2},$$

where f represents a smooth, not necessarily convex function, $\sigma > 0$ is the penalization parameter, and the norm $\|\cdot\|_{1,2}$ enforces sparsity at a group level.

To solve this general group-sparse optimization problem we introduce the Group Sparse Descent Method (GSDM). Similarly to Section 4.2, the algorithm computes the steepest descent direction of the nonsmooth problem on a group-wise basis and utilize it, together with generalized second-order information, to construct a new descent direction, resulting in a Newton-type system. Inspired on the active-set phase of the Bingham problem, for the second problem the active-set strategy is derived as an iterative interpretation of the problem's optimality condition, where the active-set for the next iteration is determined based on the angle between two consecutive iterations. The novel active-set prediction strategy is also designed to reduce the Newton-type system. Additionally, we prove that the resulting method is equivalent to a Newton-type method in a neighbourhood of a local minimum, guaranteeing fast local convergence properties. Finally, we conduct comparative computational experiments on applications to PDE-constrained optimization and non-linear regression problems to test the performance of the algorithm.

The main results of Section 4.2 were first discussed and proved in [40].

4.1 State-of-the-Art for $(\ell_{1,2})$ Norm Regularizer in Convex and Nonconvex Settings

Group sparsity aim to select a few groups of variables that serve as predictors in a large classification problem such as the group LASSO problem. Extensions to nonlin-

ear problems have also been addressed, primarily using accelerated proximal methods [83, 96]. In [83], the authors extend the accelerated proximal gradient method for general nonconvex and nonsmooth problems by introducing a monitor that satisfies the sufficient descent property. In [96], the authors propose an extrapolated proximal gradient method for nonconvex optimization problems, with cost functions composed of a continuously differentiable function and a nonsmooth one. Alternatively trust-region methods have also been proposed to tackle the minimization of nonsmooth nonconvex problems [10, 30]. In [10], the authors propose an inexact proximal trust-region method for solving problems that involve the sum of a smooth nonconvex function and a nonsmooth convex function. Also, in [30], a general trust-region method was introduced for the optimization of nonsmooth and nonconvex locally Lipschitz continuous functions. These strategies could be applied to problems involving the group sparse $\ell_{1,2}$ term.

Optimization problems governed by partial differential equations (PDE) involving group-sparsity has also been addressed, in the field of optimal control of PDEs, specific group-sparsity patterns are referred as directional sparsity (see [22, 64]). For example, in [65] the authors analyzed and solved linear-quadratic optimal control problems (both elliptic and parabolic) with a directional sparsity penalizing term by developing a semismooth Newton method.

The numerical solution of nonlinear group-sparse optimization problems is challenging due to the nonsmoothness of the $\ell_{1,2}$ norm and the non-convexity of the reduced cost function. A few second-order solution algorithms offer partial solutions to these challenges. In [84], the authors combine a semismooth Newton method with an augmented Lagrangian to solve different group LASSO problems. Furthermore, [76] proposes an inexact proximal Newton method for solving general nonconvex problems involving composite functions consisting of a twice continuously differentiable function and a convex one.

The $\ell_{1,2}$ regularizer has also been applied to dictionary learning. In [85] the research focuses on the analysis of dictionary learning using the $\ell_{1,2}$ norm to promote sparsity. The authors transform the nonconvex optimization problem into a series of one-dimensional minimization problems, allowing for efficient closed-form solutions.

Finally, for the Bingham flow problem in a pipe, in the literature we can find a range of different unregularized methods which relies in the application of the Augmented Lagrangian strategy, named as ALG1-ALG4, described in [48, 49, 52, 72] and the references therein. These methods have been extensively used to compute the solution of the Bingham flow in several contributions. For instance, in [91] the Augmented Lagrangian Method (ALM) was applied to solve multiple viscoplastic fluids in a duct. Furthermore, non-Newtonian fluids such as Herschel-Bulkley and Casson models were solved with the ALM in [71].

4.2 Group-sparse problem: Bingham flow in a pipe

This section is motivated by the analysis of the unregularized energy functional of the Bingham flow problem discussed in Chapter 3. Unlike the approach in Chapter 3, we avoid applying the C^1-regularization and instead focus directly on the nonsmooth term $g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx$. As a result, the C^1-regularization of this term used previously will not be employed here.

To further reduce the complexity arising from the presence of two nonsmooth terms in the energy functional, $g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx$ and the penalization of the divergence-free condition given by $\|\operatorname{div} \mathbf{u}\|_{L^1}$, as in Chapter 3, we will focus on the flow problem in a cylindrical pipe, where the divergence-free condition is inherently satisfied [46, Ch. IV]. Nonetheless, both non-smooth terms can be simultaneously considered. However, this is out of the scope of this thesis.

In this configuration, the domain consists of a cylinder given by a three dimensional structure whose generators are parallel to the x_3 axis in the orthonormal system of axes $x_1x_2x_3$ (see Figure 4.1). Therefore, the fluid moves just under the effect of the decay of pressure along the x_3 direction in the pipe [60, Sec. 3.1]. Thus, the velocity field is simplified to $\mathbf{u}(x_1, x_2, x_3) = (0, 0, u(x_1, x_2))$. Under these assumptions, we get that $\operatorname{div} \mathbf{u} = 0$ is automatically satisfied.

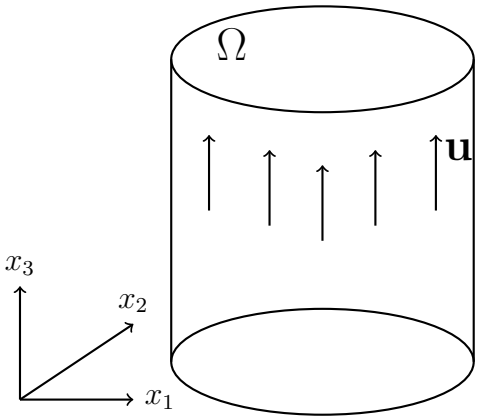


Figure 4.1: Cylindrical pipe domain.

Moreover, this simplification enables us to focus exclusively on the nonsmooth term $\int_\Omega |\mathcal{E}\mathbf{u}|\, dx$ which is simplified to $\int_\Omega |\nabla u|\, dx$. Thus, the Bingham flow problem becomes

$$\min_{u \in H_0^1(\Omega)} J(u) := \frac{\mu}{2} \int_\Omega |\nabla u|^2\, dx + \frac{g}{\sqrt{2}} \int_\Omega |\nabla u|\, dx - \int_\Omega cu\, dx,$$

where $\Omega \subset \mathbb{R}^2$ corresponds to the cross-section of a pipe and $c \in L^2(\Omega)$ represents a constant linear decay of pressure.

The unregularized approach to tackle this formulation consists in reformulating the problem as a linearly constrained minimization problem, where the constraint is defined by a new variable $\mathbf{q} = \nabla u$. Thus, the nonsmooth term becomes $\int_\Omega |\mathbf{q}|\, dx$ and we get an equivalent problem given by:

$$\min_{\substack{u\in H_0^1(\Omega),\\ \mathbf{q}\in[L^2(\Omega)]^2,\\ \mathbf{q}=\nabla u}} \widetilde{J}(u,\mathbf{q}) := \frac{\mu}{2}\int_\Omega |\nabla u|^2\, dx + \frac{g}{\sqrt{2}}\int_\Omega |\mathbf{q}|\, dx - \int_\Omega cu\, dx. \tag{4.1}$$

This structure enables the study of the nonsmooth term $\int_\Omega |\mathbf{q}|\, dx$ as a sparsity-promoting regularizer, which enforces zero values for all x in the domain Ω where the material behaves as a rigid solid, i.e., where $|\mathbf{q}(x)| = 0$. Consequently, the norm $\int_\Omega |\mathbf{q}(x)|\, dx$ promotes structured sparsity over the vector $\mathbf{q}$ in the solid-like regime. Moreover, the discretized form of the term $\int_\Omega |\mathbf{q}(x)|\, dx$ can be associated with the group-sparse norm $\|\cdot\|_{1,2}$. This connection will be revisited and explored in greater detail in the next section.

4.2.1 Augmented Lagrangian Approach and Discretization

We associate the constraint $\mathbf{q} = \nabla u$ to a Lagrange multiplier $\boldsymbol{\lambda} \in [L^2(\Omega)]^2$ (see Definition 2.3 in Chapter 2) and we define the Lagrangian functional $\mathcal{L}(u,\mathbf{q},\boldsymbol{\lambda}) : H_0^1(\Omega)\times[L^2(\Omega)]^2\times[L^2(\Omega)]^2 \to \mathbb{R}$ as follows:

$$\mathcal{L}(u,\mathbf{q},\boldsymbol{\lambda}) = \widetilde{J}(u,\mathbf{q}) + \langle \boldsymbol{\lambda}, \nabla u - \mathbf{q}\rangle_{[L^2(\Omega)]^2}. \tag{4.2}$$

Moreover, the augmented Lagrangian functional denoted by $\mathcal{L}_\rho(u,\mathbf{q},\boldsymbol{\lambda})$, for $\rho > 0$, is given by

$$\mathcal{L}_\rho(u,\mathbf{q},\boldsymbol{\lambda}) = \widetilde{J}(u,\mathbf{q}) + (\boldsymbol{\lambda}, \nabla u - \mathbf{q})_{[L^2(\Omega)]^2} + \frac{\rho}{2}\|\nabla u - \mathbf{q}\|^2_{[L^2(\Omega)]^2}. \tag{4.3}$$

In the classical Augmented Lagrangian Method (ALM), as described in [57], solving the minimization problem (4.1) is equivalent to finding the saddle point of the augmented Lagrangian functional $\mathcal{L}_\rho(u,\mathbf{q},\boldsymbol{\lambda})$. Thus, there exists a saddle point of $\mathcal{L}(u,\mathbf{q},\boldsymbol{\lambda})$ satisfying the constraint $\mathbf{q} = \nabla u$ which is also a saddle point of the augmented Lagrangian $\mathcal{L}_\rho(u,\mathbf{q},\boldsymbol{\lambda})$ [49].

We discretize the Augmented Lagrangian problem by using the Galerkin Finite Element Method. For the velocity field u, linear polynomials are employed, while $\mathbf{q}$ and $\boldsymbol{\lambda}$ are approximated using piecewise constant functions (see [91]). The domain Ω is assumed to be an open subset of $\mathbb{R}^2$. Let Ω_h denote a regular triangulation of Ω with

conforming elements such that $\bar{\Omega}_h = \cup_{T\in\Omega_h} T$. Taking this into account, we define

$$V_{1_h} := \{v_h \in C(\bar{\Omega}) : v_h|_T \in P^1, \ \ \forall\, T \in \Omega_h\},$$

and

$$V_{2_h} := \{\mathbf{w}_h = (w_{1_h}, w_{2_h}) \in [L^2(\Omega)]^2 : w_{1_h}|_T, w_{2_h}|_T \in P^0, \ \ \forall\, T \in \Omega_h\},$$

where P^1 and P^0 are the spaces of continuous piecewise linear functions and piecewise constant functions, respectively defined on Ω_h. Moreover, the spaces $V^0_{1_h} = H^1_0(\Omega) \cap V_{1_h}$ and V_{2_h} are the finite-dimensional spaces associated with Ω_h.

Considering the previous analysis, the finite element approximation of (4.3) is formulated as follows:

$$\mathcal{L}_\rho(u_h, \mathbf{q}_h, \boldsymbol{\lambda}_h) = \widetilde{J}_h(u_h, \mathbf{q}_h) + (\boldsymbol{\lambda}_h, \nabla_h u_h - \mathbf{q}_h)_h + \frac{\rho}{2}\|\nabla_h u_h - \mathbf{q}_h\|^2_h, \tag{4.4}$$

where,

$$\widetilde{J}_h(u_h, \mathbf{q}_h) := \frac{\mu}{2}\int_{\Omega_h} |\nabla u_h|^2\, dx + \frac{g}{\sqrt{2}}\int_{\Omega_h} |\mathbf{q}_h|\, dx - \int_{\Omega_h} c_h u_h\, dx \tag{4.5}$$

represents the finite-element approximation of the functional $\widetilde{J}$. Moreover ∇_h corresponds to the finite-element approximation of the gradient operator. Also, $(\cdot,\cdot)_h$ and $\|\cdot\|^2_h$ denote the finite-element approximations of the inner product and the norm in $[L^2(\Omega)]^2$, respectively.

The framework of the augmented Lagrangian method allows us to interpret the constraint term $\int_{\Omega_h} |\mathbf{q}_h|\, dx$ as a sparsity-promoting regularizer, enforcing zero values where the material behaves as a rigid solid, i.e., at the triangles $T \in \Omega_h$ where $|\mathbf{q}_h| = |\nabla_h u_h| = \left|\left(\frac{\partial u_h}{\partial x_1}, \frac{\partial u_h}{\partial x_2}\right)\right| = 0$. As a result, the norm $\int_{\Omega_h} |\mathbf{q}|\, dx$ is expressed as:

$$\int_{\Omega_h} |\mathbf{q}_h|\, dx = \int_{\Omega_h} \left|\left(\frac{\partial u_h}{\partial x_1}, \frac{\partial u_h}{\partial x_2}\right)\right| dx = \int_{\Omega_h} |(q_{1_h}, q_{2_h})|\, dx = \int_{\Omega_h} \sqrt{q^2_{1_h} + q^2_{2_h}}\, dx.$$

Specifically, the pairs (q_{1_h}, q_{2_h}) can be analyzed as "groups" and categorized based on whether they exhibit sparsity. By denoting p the total number of the elements in the triangulation Ω_h, the term $\int_{\Omega_h} |\mathbf{q}_h|\, dx$ can be approximated by

$$\sum_{i=1}^{p} |\mathbf{q}_{h_i}|.$$

Where $|\mathbf{q}_{h_i}|$ is the value of $\sqrt{q^2_{1_{h_i}} + q^2_{2_{h_i}}}$ over each i-th triangle of Ω_h. Thus, $\sum_{i=1}^{p} |\mathbf{q}_{h_i}|$ corresponds to the discrete group-sparse norm $\|\cdot\|_{1,2}$.

We review the well know ALG1 method from the Augmented Lagrangian framework [49], applied to problem (4.4) :

Algorithm 3: ALG1

Set $k = 0$, initialize $\boldsymbol{\lambda}_h^0$;

while *stopping criterion is not satisfied* **do**

with $\boldsymbol{\lambda}_h^k$ known, compute

$$(u_h^k, \mathbf{q}_h^k) = \underset{(u_h,\mathbf{q}_h)\in V_{1_h}^0 \times V_{2_h}}{\arg\min} L((u_h, \mathbf{q}_h)) = \widetilde{J}_h(u_h, \mathbf{q}_h) + (\boldsymbol{\lambda}_h, \nabla_h u_h - \mathbf{q}_h)_h + \frac{\rho}{2}\|\nabla_h u_h - \mathbf{q}_h\|_h^2 \quad \textbf{(IOP)}$$

compute a step size $t^k > 0$;

update $\boldsymbol{\lambda}_h^{k+1} = \boldsymbol{\lambda}_h^k + t^k(\nabla_h u_h^k - \mathbf{q}_h^k),$;

set $k \leftarrow k + 1$

end

We will refer to the inner optimization problem in ALG 1 as (**IOP**). The convergence of ALG 1 is particularly slow in the rigid zones of the fluid [49, Sec. 6.3.1], making the algorithm computationally expensive [111].

An alternative to ALG 1 is ALG 2, also known as the Alternating Direction Method of Multipliers (ADMM). This method splits the solution of (**IOP**) by performing a minimization in u_h^k followed by a minimization in $\mathbf{q}_h^k$. This approach is related to the Douglas–Rachford splitting algorithm and can be viewed as solving (**IOP**) inexactly. This approach can be traced back to [52]. In the literature, ALG 2 is regarded as the standard nonsmooth approach for simulating viscoplastic fluid flows [58].

However, based on the strategies developed in Chapter 3, including the use of the steepest descent direction and its acceleration through generalized second-order information, we aim to apply these tools to enhance the solution of problem (**IOP**), addressing u_h^k and $\mathbf{q}_h^k$ jointly. We exploit the group structure of the constraint $\mathbf{q}_h$ to identify the rigid zones at each iterate. This phase will be referred as the active set prediction phase.

4.2.2 Steepest descent direction

To solve (**IOP**) we propose implementing a descent algorithm. The function $L((u_h, \mathbf{q}_h))$ in problem (**IOP**) is not differentiable due to the term $g \int_{\Omega_h} |\mathbf{q}_h|\, dx$ in $\widetilde{J}(u_h, \mathbf{q}_h)$. Nevertheless, $L((u_h, \mathbf{q}_h))$ is directionally differentiable because the Euclidean norm $|\mathbf{q}_h|$ itself is directionally differentiable. Let us denote $h(\mathbf{q}_h) = |\mathbf{q}_h|$. The directional derivative

of h at $\mathbf{q}_h$ in the direction $\mathbf{d_q}$, is given by:

$$h'(\mathbf{q}_h, \mathbf{d_q}) = \begin{cases} \dfrac{\langle \mathbf{q}_h, \mathbf{d_q} \rangle}{|\mathbf{q}_h|}, & \text{if } |\mathbf{q}_h| \neq \mathbf{0}, \\ |\mathbf{d_q}|, & \text{otherwise.} \end{cases} \tag{4.6}$$

Then, the directional derivative of $L(u_h, \mathbf{q}_h)$ at $(u_h, \mathbf{q}_h)$ in the direction $(d_u, \mathbf{d_q})$ is denoted by $L^{'}((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q}))$ and given by:

$$L^{'}((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q})) = \widetilde{J}'_h((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q})) + (\boldsymbol{\lambda}_h, \nabla_h d_u - \mathbf{d_q}) + \rho(\nabla u_h - \mathbf{q}_h, \nabla_h d_u - \mathbf{d_q}), \tag{4.7}$$

where $\widetilde{J}'_h((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q}))$ denotes the directional derivative of $\widetilde{J}_h$ at $(u_h, \mathbf{q}_h)$ in the direction $(d_u, \mathbf{d_q})$. Moreover, $\widetilde{J}'_h((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q}))$ is given by:

$$\begin{aligned} \widetilde{J}'_h((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q})) &= \mu \int_{\Omega_h} \nabla_h u_h \cdot \nabla_h d_u \, dx - \int_{\Omega_h} c_h d_u \, dx + \int_{\Omega_h} h'(\mathbf{q}_h, \mathbf{d_q}) \, dx \\ &= \mu \int_{\Omega_h} -\Delta_h u_h \cdot d_u \, dx - \int_{\Omega_h} c_h d_u \, dx + \frac{g}{\sqrt{2}} \int_{\Omega_h} h'(\mathbf{q}_h, \mathbf{d_q}) \, dx \\ &= (-\mu \Delta_h u_h \cdot d_u)_h - (c_h, d_u)_h + \int_{\Omega_h} h'(\mathbf{q}_h, \mathbf{d_q}) \, dx. \end{aligned} \tag{4.8}$$

The terms in (4.7) can be reorganized to rewrite the directional derivative of $L(u_h, \mathbf{q}_h)$, given by $L'((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q}))$, as an additive separable function in the two directions. Specifically, (4.7) can be expressed as the sum of two functions, each depending separately on d_u and $\mathbf{d_q}$. Therefore, we introduce the functions $P_h(u_h, \mathbf{q}_h) : V^0_{1_h} \to \mathbb{R}$ and $Q_h(u_h, \mathbf{q}_h) : V_{2_h} \to \mathbb{R}$ such that:

$$L^{'}((u_h, \mathbf{q}_h), (d_u, \mathbf{d_q})) = (P_h(u_h, \mathbf{q}_h), d_u) + (Q_h(u_h, \mathbf{q}_h), \mathbf{d_q}) + \frac{g}{\sqrt{2}} \int_{\Omega_h} h'(\mathbf{q}_h, \mathbf{d_q}) \, dx, \tag{4.9}$$

where

$$(P_h(u_h, \mathbf{q}_h), d_u) = (-\Delta_h u_h \cdot d_u)_h - (c_h, d_u)_h - (\mathrm{div}_h \boldsymbol{\lambda}_h, d_u)_h - \rho(\mathrm{div}_h(\nabla u_h - \mathbf{q}_h), d_u)_h \tag{4.10}$$

and

$$(Q_h(u_h, \mathbf{q}_h), \mathbf{d_q}) = -(\boldsymbol{\lambda}_h, \mathbf{d_q})_h - \rho(\nabla_h u_h - \mathbf{q}_h, \mathbf{d_q})_h.$$

For simplicity, in the following discussion, we omit the subscript h when referring to discrete inner products.

After computing the directional derivative of the function in problem (**IOP**), we can determine the steepest descent direction. From Chapter 2, Proposition 2.2, the steepest descent direction is defined as the direction that minimizes the directional derivative over the unit ball. Accordingly, obtaining this direction requires solving the

following problem:

$$(\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q}) = \underset{\|(\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q})\| \leq 1}{\arg\min}\ L^{'}((u_h, \mathbf{q}_h), (d_u, \mathbf{d}_\mathbf{q})). \tag{4.11}$$

Where $\|(\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q})\|$ corresponds to the Euclidean norm of the joint vector $(\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q})$.

Given the inequalities:

$$\|\hat{d}_u\| \leq \|(\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q})\| \leq 1,$$

$$\|\hat{\mathbf{d}}_\mathbf{q}\| \leq \|(\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q})\| \leq 1,$$

and the fact that $L^{'}((u_h, \mathbf{q}_h), (d_u, \mathbf{d}_\mathbf{q}))$, as defined in (4.9), is an additive separable function in the direction, we can split problem (4.11) in two problems and the steepest descent directions $\hat{d}_u$ and $\hat{\mathbf{d}}_\mathbf{q}$ can be determined independently by solving:

$$\hat{d}_u = \underset{\|d_u\| \leq 1}{\arg\min}\ (P_h(u_h, \mathbf{q}_h), d_u) \qquad \text{and} \tag{4.12}$$

$$\hat{\mathbf{d}}_\mathbf{q} = \underset{\|\mathbf{d}_\mathbf{q}\| \leq 1}{\arg\min}\ (Q_h(u_h, \mathbf{q}_h), \mathbf{d}_\mathbf{q}) + \frac{g}{\sqrt{2}} \int_{\Omega_h} h'(\mathbf{q}_h, \mathbf{d}_\mathbf{q})\, dx. \tag{4.13}$$

The norms chosen in problems (4.12) and (4.13) influence the resulting solutions. In our case, the groupwise structure of the term $\int_{\Omega_h} |\mathbf{q}(x)|\, dx$, interpreted as the group-sparse norm $\|\cdot\|_{1,2}$, motivates us to replace the norm $\|\mathbf{d}_\mathbf{q}\|$ with the dual norm $\|\mathbf{d}_\mathbf{q}\|_{\infty,2}$. Thus the steepest descent direction in (4.13) is considered in the following sense:

$$\hat{\mathbf{d}}_\mathbf{q} = \underset{\|\mathbf{d}_\mathbf{q}\|_{\infty,2} \leq 1}{\arg\min}\ (Q_h(u_h, \mathbf{q}_h), \mathbf{d}_\mathbf{q}) + \frac{g}{\sqrt{2}} \int_{\Omega_h} h'(\mathbf{q}_h, \mathbf{d}_\mathbf{q})\, dx. \tag{4.14}$$

Recalling that $\int_{\Omega_h} h'(\mathbf{q}_h, \mathbf{d}_\mathbf{q})$ can be approximated by $\sum_{i=1}^{p} h'(\mathbf{q}_{h_i}, \mathbf{d}_{\mathbf{q}_i})$ and thanks to the definition of the norm $\|\mathbf{d}_\mathbf{q}\|_{\infty,2} = \max_{i=1,\cdots,p} |\mathbf{d}_{\mathbf{q}_i}|$, we can derive p independent subproblems from (4.14):

$$\hat{\mathbf{d}}_{\mathbf{q}_i} = \underset{|\mathbf{d}_{\mathbf{q}_i}| \leq 1}{\arg\min}\ (Q_{h_i}(u_h, \mathbf{q}_h), \mathbf{d}_{\mathbf{q}_i}) + \frac{g}{\sqrt{2}} h'(\mathbf{q}_{h_i}, \mathbf{d}_{\mathbf{q}_i}), \text{ for all } i = 1, \cdots, p. \tag{4.15}$$

where $Q_{h_i}(u_h, \mathbf{q}_h)$ corresponds to the i-th component of the vector $Q_h(u_h, \mathbf{q}_h)$.

Theorem 4.1. *The unique solution to problem (4.15) is given by*

$$\hat{\mathbf{d}}_{\mathbf{q}_i} = \begin{cases} -\dfrac{Q_{h_i}(u_h, \mathbf{q}_h) + \frac{g}{\sqrt{2}} \frac{\mathbf{q}_{h_i}}{|\mathbf{q}_{h_i}|}}{\left| Q_{h_i}(u_h, \mathbf{q}_h) + \frac{g}{\sqrt{2}} \frac{\mathbf{q}_{h_i}}{|\mathbf{q}_{h_i}|} \right|}, & if\ \ |\mathbf{q}_{h_i}| \neq 0, \\ \mathbf{0}, & if\ \ |\mathbf{q}_{h_i}| = 0\ \ and\ \ |Q_{h_i}(u_h, \mathbf{q}_h)| \leq \frac{g}{\sqrt{2}}, \\ -\dfrac{Q_{h_i}(u_h, \mathbf{q}_h)}{|Q_{h_i}(u_h, \mathbf{q}_h)|}, & if\ \ |\mathbf{q}_{h_i}| = 0\ \ and\ \ |Q_{h_i}(u_h, \mathbf{q}_h)| > \frac{g}{\sqrt{2}}, \end{cases} \tag{4.16}$$

for all $i = 1, \cdots, p$.

Proof. Using the definition of $h'(\mathbf{q}_{h_i}, \mathbf{d}_{\mathbf{q}_i})$ provided in (4.6), we compute the steepest descent direction $\hat{\mathbf{d}}_{\mathbf{q}_i}$ by analyzing whether $|\mathbf{q}_{h_i}|$ exhibits sparsity or not, i.e., whether $|\mathbf{q}_{h_i}| = 0$ or $|\mathbf{q}_{h_i}| \neq 0$. Accordingly, from (4.15) we obtain:

if $|\mathbf{q}_{h_i}| \neq 0$, by using the Cauchy-Schwarz inequality, we get that

$$\begin{aligned}(Q_{h_i}(u_h, \mathbf{q}_h), \mathbf{d}_{\mathbf{q}_i}) + g h'(\mathbf{q}_{h_i}, \mathbf{d}_{\mathbf{q}_i}) = & \ (Q_{h_i}(u_h, \mathbf{q}_h) + \tfrac{g}{\sqrt{2}} \tfrac{\mathbf{q}_{h_i}}{|\mathbf{q}_{h_i}|}, \mathbf{d}_{\mathbf{q}_i}) \\ \geq & \ -\left|Q_{h_i}(u_h, \mathbf{q}_h) + \tfrac{g}{\sqrt{2}} \tfrac{\mathbf{q}_{h_i}}{|\mathbf{q}_{h_i}|}\right| \left|\mathbf{d}_{\mathbf{q}_i}\right| \\ \geq & \ -\left|Q_{h_i}(u_h, \mathbf{q}_h) + \tfrac{g}{\sqrt{2}} \tfrac{\mathbf{q}_{h_i}}{|\mathbf{q}_{h_i}|}\right|,\end{aligned} \tag{4.17}$$

where the last inequality in (4.17) is given because of the restriction $\left|\mathbf{d}_{\mathbf{q}_i}\right| \leq 1$. Therefore, the solution to (4.15) is attained at:

$$\hat{\mathbf{d}}_{\mathbf{q}_i} = -\frac{Q_{h_i}(u_h, \mathbf{q}_h) + \frac{g}{\sqrt{2}} \frac{\mathbf{q}_{h_i}}{|\mathbf{q}_{h_i}|}}{\left|Q_{h_i}(u_h, \mathbf{q}_h) + \frac{g}{\sqrt{2}} \frac{\mathbf{q}_{h_i}}{|\mathbf{q}_{h_i}|}\right|}.$$

On the other hand, if $|\mathbf{q}_{h_i}| = 0$, it follows from problem (4.15) and relation (4.6) that

$$(Q_{h_i}(u_h, \mathbf{q}_h), \mathbf{d}_{\mathbf{q}_i}) + \frac{g}{\sqrt{2}} h'(\mathbf{q}_{h_i}, \mathbf{d}_{\mathbf{q}_i}) = \ (Q_{h_i}(u_h, \mathbf{q}_h), \mathbf{d}_{\mathbf{q}_i}) + \tfrac{g}{\sqrt{2}} |\mathbf{d}_{\mathbf{q}_i}|.$$

Let us suppose that $(Q_{h_i}(u_h, \mathbf{q}_h), \mathbf{d}_{\mathbf{q}_i}) + \frac{g}{\sqrt{2}} |\mathbf{d}_{\mathbf{q}_i}| \geq 0$ then, the minimum is attained at zero, i.e., $\hat{\mathbf{d}}_{\mathbf{q}_i} = \mathbf{0}$. On the other hand, if $(Q_{h_i}(u_h, \mathbf{q}_h), \mathbf{d}_{\mathbf{q}_i}) + \frac{g}{\sqrt{2}} |\mathbf{d}_{\mathbf{q}_i}|$ is negative, from $\left|\mathbf{d}_{\mathbf{q}_i}\right| \leq 1$, we have

$$\begin{aligned}0 > (Q_{h_i}(u_h, \mathbf{q}_h), \mathbf{d}_{\mathbf{q}_i}) + \tfrac{g}{\sqrt{2}} |\mathbf{d}_{\mathbf{q}_i}| \geq & \ -|Q_{h_i}(u_h, \mathbf{q}_h)||\mathbf{d}_{\mathbf{q}_i}| + \frac{g}{\sqrt{2}} |\mathbf{d}_{\mathbf{q}_i}| \\ \geq & \ -(|Q_{h_i}(u_h, \mathbf{q}_h)| - \frac{g}{\sqrt{2}}) |\mathbf{d}_{\mathbf{q}_i}| \\ \geq & \ -(|Q_{h_i}(u_h, \mathbf{q}_h)| - \frac{g}{\sqrt{2}}).\end{aligned}$$

Thus, the lower bound is attained at $\hat{\mathbf{d}}_{\mathbf{q}_i} = -\dfrac{Q_{h_i}(u_h, \mathbf{q}_h)}{|Q_{h_i}(u_h, \mathbf{q}_h)|}$. Moreover, this case holds whenever $|Q_{h_i}(u_h, \mathbf{q}_h)| > \dfrac{g}{\sqrt{2}}$.

By collecting all the cases, we find that the solution to problem (4.15) corresponds to the steepest descent direction $\hat{\mathbf{d}}_{\mathbf{q}_i}$ for all $i = 1, \ldots, p$, given by:

$$\hat{\mathbf{d}}_{\mathbf{q}i} = \begin{cases} -\dfrac{Q_{h_i}(u_h, \mathbf{q}_h) + \frac{g}{\sqrt{2}} \frac{\mathbf{q}_{h_i}}{|\mathbf{d}_{\mathbf{q}_i}|}}{\left| Q_{h_i}(u_h, \mathbf{q}_h) + \frac{g}{\sqrt{2}} \frac{\mathbf{q}_{h_i}}{|\mathbf{d}_{\mathbf{q}_i}|} \right|}, & if \ \ |\mathbf{q}_{h_i}| \neq 0 \\ \mathbf{0}, & if \ \ |\mathbf{q}_{h_i}| = 0 \text{ and } \ |Q_{h_i}(u_h, \mathbf{q}_h)| \leq \frac{g}{\sqrt{2}} \\ -\dfrac{Q_{h_i}(u_h, \mathbf{q}_h)}{|Q_{h_i}(u_h, \mathbf{q}_h)|}, & if \ \ |\mathbf{q}_{h_i}| = 0 \text{ and } \ |Q_{h_i}(u_h, \mathbf{q}_h)| > \frac{g}{\sqrt{2}}. \end{cases}$$

□

It remains to compute the steepest descent direction $\hat{d}_u$ from problem (4.12).

Theorem 4.2. *The steepest descent direction $\hat{d}_u$ from problem (4.12) is given by:*

$$\hat{d}_u = -\frac{P_h(u_h, \mathbf{q}_h)}{|P_h(u_h, \mathbf{q}_h)|}. \tag{4.18}$$

Proof. Since $L((u_h, \mathbf{q}_h))$ is differentiable with respect to u_h, the directional derivative in the direction d_u, given in (4.10), coincides with the partial derivative of $L((u_h, \mathbf{q}_h))$ with respect to u_h. Thus, we deduce that the steepest descent direction with respect to u_h is given by (4.18). □

Consequently, from Theorems 4.1 and 4.2 we have computed the steepest descent direction $(\hat{d}_u, \hat{\mathbf{d}}_{\mathbf{q}})$ with $\hat{d}_u$ given in (4.18) and $\hat{\mathbf{d}}_{\mathbf{q}}$ given componentwise in (4.16).

4.2.3 Second-order Information

Following the methodology derived for the regularized optimization problem in Chapter 3, our strategy incorporates curvature information provided by generalized second-order derivatives. Thus, to solve de inner optimization problem (**IOP**) of ALG 1, we propose to modify the steepest descent direction $(\hat{d}_u, \hat{\mathbf{d}}_{\mathbf{q}})$ by incorporating second order information of the function

$$L((u_h, \mathbf{q}_h)) = \widetilde{J}_h(u_h, \mathbf{q}_h) + (\boldsymbol{\lambda}_h, \nabla_h u_h - \mathbf{q}_h) + \frac{\rho}{2} \|\nabla u_h - \mathbf{q}_h\|^2.$$

.

Note that $\widetilde{J}_h(u_h, \mathbf{q}_h)$, given in (4.5), involves the non differentiable term $g \int_{\Omega_h} |\mathbf{q}_h| \, dx \approx g \sum_{i=1}^{p} |\mathbf{q}_{h_i}| \, dx$. In this section, unlike Chapter 3, we will not replace the Euclidean norm $|\cdot|$ with its regularized formulation. Nevertheless, we will use the regularized version of the norm to obtain a generalized second order derivative to enrich the curvature.

Since $h(\mathbf{q}_{h_i}) = |\mathbf{q}_{h_i}|$, in this Chapter the Huber regularization of the Euclidean

norm is denoted by the function h_γ, with approximation parameter $\gamma > 0$, such that:

$$h_\gamma(\mathbf{q}_{h_i}) = \begin{cases} |\mathbf{q}_{h_i}| - \frac{1}{2\gamma}, & if\ |\mathbf{q}_{h_i}| > \frac{1}{\gamma}, \\ \frac{\gamma}{2}|\mathbf{q}_{h_i}|^2, & if\ |\mathbf{q}_{h_i}| \leq \frac{1}{\gamma}, \end{cases}$$

for all $i = 1, \cdots, p$. Function h_γ is differentiable and its gradient is given by:

$$\nabla h_\gamma(\mathbf{q}_{h_i}) = \frac{\gamma \mathbf{q}_{h_i}}{\max(1, \gamma|\mathbf{q}_{h_i}|)}, \ \ \forall i = 1, \cdots, p.$$

This gradient is not differentiable because of the max function. However, it is locally Lipschitz continuous and directionally differentiable. Thus, by Remark 2.2, it is Bouligand differentiable. Additionally, since the max function is semismooth (as shown in Example 2.2), the function $\nabla h_\gamma(\mathbf{u}_i)$ is also semismooth, due to the composition property of semismooth functions established in Proposition 2.4.

Using the rules of the Bouligand subdifferentials, we can compute the second-order generalized derivative of h_γ, denoted by $\Gamma(\mathbf{q}_{h_i})$ and given by

$$\Gamma(\mathbf{q}_{h_i}) = \begin{cases} \dfrac{1}{|\mathbf{q}_{h_i}|} I - \dfrac{\mathbf{q}_{h_i}\mathbf{q}_{h_i}^\top}{|\mathbf{q}_{h_i}|^3}, & \text{if } |\mathbf{q}_{h_i}| \geq \frac{1}{\gamma}, \\ \gamma I, & \text{if } |\mathbf{q}_{h_i}| < \frac{1}{\gamma}, \end{cases} \tag{4.19}$$

for all $i = 1, \ldots, p$.

The second-order information incorporates both the generalized derivative $\Gamma(\mathbf{q}_h)$ and the second Fréchet derivative of the remaining differentiable terms in $L((u_h, \mathbf{q}_h))$. The bilinear form in $V_{1_h}^0 \times V_{2_h}$ containing second-order information is denoted by $\mathcal{H}(u_h, \mathbf{q}_h)(\cdot, \cdot)$ such that, for every $\mathbf{v}_h = (v_{h_u}, \mathbf{v}_{\mathbf{h_q}})$ and $\mathbf{z}_h = (z_{h_u}, \mathbf{z}_{\mathbf{h_q}}) \in V_{1_h}^0 \times V_{2_h}$, it is given by

$$\begin{aligned} \mathcal{H}(u_h, \mathbf{q}_h)(\mathbf{z}_h, \mathbf{v}_h) = \ & (\mu + \rho)(\nabla_h z_{hu} \cdot \nabla_h v_{hu}) - \rho(\, \mathbf{z}_{\mathbf{h_q}}, \nabla_h v_{hu}) - \rho(\, \nabla_h z_{hu}, \mathbf{v}_{\mathbf{h_q}}) \\ & + \rho(\, \mathbf{z}_{\mathbf{h_q}}, \mathbf{v}_{\mathbf{h_q}}) + \frac{g}{\sqrt{2}} \int_{\Omega_h} \Gamma(\mathbf{q}_h)(\mathbf{v}_{\mathbf{h_q}}, \mathbf{z}_{\mathbf{h_q}}) \\ & = ((\mu + \rho)\Delta_h z_{hu} - \rho\ \operatorname{div}_h \mathbf{z}_{\mathbf{h_q}}, v_{hu}) - \rho(\, \nabla_h z_{hu} - \mathbf{z}_{\mathbf{h_q}}, \mathbf{v}_{\mathbf{h_q}}) \\ & + \frac{g}{\sqrt{2}} \int_{\Omega_h} \Gamma(\mathbf{q}_h)(\mathbf{v}_{\mathbf{h_q}}, \mathbf{z}_{\mathbf{h_q}})\, dx. \end{aligned} \tag{4.20}$$

Next, in order to obtain the Newton-type direction $\mathbf{w}_h = (w_{h_u}, \mathbf{w}_{\mathbf{h_q}})$ we have to solve the following variational problem:

$$\mathcal{H}(u_h, \mathbf{q}_h)(\mathbf{w}_h, \mathbf{v}_h) = ((\hat{d}_u, \hat{\mathbf{d}}_{\mathbf{q}}), \mathbf{v}_h), \qquad \forall\ \mathbf{v}_h \in V_{1_h}^0 \times V_{2_h}, \tag{4.21}$$

where

$$\begin{aligned}((\hat{d}_u, \hat{\mathbf{d}}_{\mathbf{q}}), \mathbf{v}_h) &= ((\hat{d}_u, \hat{\mathbf{d}}_{\mathbf{q}}), (v_{h_u}, \mathbf{v}_{\mathbf{h_q}})) \\ &= (\hat{d}_u, v_{h_u}) + (\hat{\mathbf{d}}_{\mathbf{q}}, \mathbf{v}_{\mathbf{h_q}}) \\ &= (P_h(u_h, \mathbf{q}_h), v_{hu}) + (\hat{\mathbf{d}}_{\mathbf{q}}, \mathbf{v}_{\mathbf{h_q}}).\end{aligned} \tag{4.22}$$

Using the constructed Newton-type direction, we can address the inner minimization problem (**IOP**) in ALG 1. Solving this optimization problem necessitates the implementation of a suitable line search strategy.

4.2.4 Line-search Strategy

In the inner optimization loop (**IOP**) of Algorithm 1, the iterations will be indexed by j. Thus, given a current iteration $(u_h, \mathbf{q}_h)^j$ and the Newton-type search direction $\mathbf{w}_h = (w_{h_u}, \mathbf{w}_{\mathbf{h_q}})^j \in V_{1_h}^0 \times V_{2_h}$, the goal of the line search is to determine a step size $s^j > 0$ that reduces the objective function along the line $(u_h, \mathbf{q}_h)^j + s^j(w_{h_u}, \mathbf{w}_{\mathbf{h_q}})^j$. To achieve this, we employ a generalized Armijo condition specifically designed for nonsmooth convex functions introduced in [122]. This condition extends the traditional Wolfe conditions through a subgradient reformulation.

This subgradient-based reformulation for an arbitrary non-smooth convex function J at iterate x^j, in the direction w^j, is expressed as:

$$J(x^j + s^j w^j) \leq J(x^j) + c_0 s^j \sup_{g \in \partial J(x^j)} (w^j, g), \tag{4.23}$$

where $c_0 > 0$ is a fixed constant.

Consequently, since the subdifferential of a convex function can also be characterized by the directional derivative (see Definition 2.2 in Chapter 2) we can rewrite (4.23) as:

$$J(x^j + s^j w^j) \leq J(x^j) + c_0 s^j J'(x^j, w^j). \tag{4.24}$$

In particular, this generalization of the Armijo condition (4.24), applied to the functional $L((u_h, \mathbf{q}_h))$, is expressed as:

$$L(u_h, \mathbf{q}_h)^{j+1} \leq L((u_h, \mathbf{q}_h))^j + c_0 s^j L'((u_h, \mathbf{q}_h)^j, (w_{h_u}, \mathbf{w}_{\mathbf{h_q}})^j). \tag{4.25}$$

With the stepsize s^j and the Newton-type direction $(w_{h_u}^j, \mathbf{w}_{\mathbf{h_q}}{}^j)$, the iteration for solving problem (**IOP**) is updated as follows:

$$(u_h, \mathbf{q}_h)^{j+1} = (u_h, \mathbf{q}_h)^j + s^j(w_{h_u}, \mathbf{w}_{\mathbf{h_q}})^j. \tag{4.26}$$

Then, after finding the minimizer $(u_h, \mathbf{q}_h)^k$ of problem (**IOP**), we can apply recursively ALG 1 and obtain the solution of the augmented Lagrangian formulation (4.4).

4.2.5 Active-set Prediction Phase

Convergence of ALG 1 is particularly slow in the solid-like regime [49, Sec. 6.3.1]. To enhace the identification of the rigid zones, we propose an active-set prediction phase to *a priori* identify the indices $i = 1, \cdots, p$, where $|\mathbf{q}_{h_i}|$ is sparse. To achieve this purpose, we recall some projection methods, such as the orthant-based methods proposed in [1] and [21]. These methods perform orthogonal projections onto a predefined orthant face to determine the variables that are active (sparse) in the solution. These methods set to zero any coordinate that transitions from positive to negative, or vice versa, during the iteration. In [39], a second-order algorithm was introduced to cope with non-convex problems with point-wise sparsity induced by the ℓ_1-norm. It utilizes an orthogonal projection step onto the orthant face of inactive components to prevent coordinates from changing signs. We extend this strategy to the term $\sum_{i=1}^{p} |\mathbf{q}_{h_i}|$ to determine, within the inner optimization problem (**IOP**), whether $\mathbf{q}_{h_i}^j$ exhibits sparsity at the i-th triangle during the j-th iteration.

We motivate the active set prediction phase by providing a geometrical interpretation. To elaborate, let us consider an i-th triagle at the j-th iteration where $|\mathbf{q}_{h_i}^j| > 0$. Given the search direction ${\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j$ computed by solving system (4.21), we compare two consecutive iterations of vector $\mathbf{q}_{h_i}$. Then, we have that:

$$\begin{aligned} (\mathbf{q}_{h_i}^{j+1}, \mathbf{q}_{h_i}^j) &= (\mathbf{q}_{h_i}^j + s^j {\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j, \mathbf{q}_{h_i}^j) \\ &= |\mathbf{q}_{h_i}^j|^2 + s^j ({\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j, \mathbf{q}_{h_i}^j). \end{aligned} \tag{4.27}$$

If $(\mathbf{q}_{h_i}^{j+1}, \mathbf{q}_{h_i}^j) \geq 0$ we have that the two consecutive iterations $\mathbf{q}_{h_i}^j$ and $\mathbf{q}_{h_i}^{j+1}$ form a right or acute angle. Conversely, given that $|\mathbf{q}_{h_i}^j|^2 > 0$, if $(\mathbf{q}_{h_i}^{j+1}, \mathbf{q}_{h_i}^j) < 0$, then $({\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j, \mathbf{q}_{h_i}^j) < 0$. In this case, the vectors $\mathbf{q}_{h_i}^j$ and ${\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j$ form an obtuse angle, implying that the descent direction ${\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j$ differs significantly from $\mathbf{q}_{h_i}^j$ and points in a different direction. Intuitively, this change in the direction may happen when a group is close to becoming sparse and the descent direction promotes zig-zagging effects.

To avoid this behaviour, we propose the following active-set prediction phase: we set the updated solution $\mathbf{q}_{h_i}^{j+1}$ to exactly zero whenever $(\mathbf{q}_{h_i}^j + s^j {\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j, \mathbf{q}_{h_i}^j) < 0$. Hence, we have that

$$\mathbf{q}_{h_i}^{j+1} = \begin{cases} \mathbf{0}, & \text{if } (\mathbf{q}_{h_i}^j + s^j {\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j, \mathbf{q}_{h_i}^j) < 0 \\ \mathbf{q}_{h_i}^j + s^j {\mathbf{w}_{\mathbf{h}_{\mathbf{q}_\mathbf{i}}}}^j, & \text{otherwise.} \end{cases}$$

4.2.6 Second Order Optimization Algorithm and Numerical experiments

We present a summary of the strategies previously developed for solving the optimization problem (**IOP**), structured in Algorithm 4.

Moreover, to solve the augmented Lagrangian formulation (4.4), we conduct numerical experiments using ALG1 in combination with Algorithm 4 as the inner optimization method for problem (**IOP**). Additionally, we compare the performance of this approach with the ALG2 method, which is widely regarded as the standard nonsmooth approach for simulating viscoplastic fluid flows. The results of this comparison are presented in Table 4.1 and Figure 4.2.

Algorithm 4: Group-sparse algorithm for ALG1

Set $j = 0$;

while *stopping criterion is not satisfied* **do**

compute the steepest descent direction $(\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q})$ given in (4.18) and (4.16) ;

compute the Newton-type descent direction $\mathbf{w}_h^j = (w_{h_u}, \mathbf{w}_{\mathbf{h_q}})^j \in V_{1_h}^0 \times V_{2_h}$ by solving the system:

$$\mathcal{H}(u_h, \mathbf{q}_h)(\mathbf{w}_h, \mathbf{v}_h) = ((\hat{d}_u, \hat{\mathbf{d}}_\mathbf{q}), \mathbf{v}_h), \qquad \forall\ \mathbf{v}_h \in V_{1_h}^0 \times V_{2_h},$$

execute a line-search such that s^j satisfies the generalized condition (4.25);

update

$$\mathbf{q}_{h_i}^{j+1} = \begin{cases} \mathbf{0}, & \text{if } (\mathbf{q}_{h_i}^j + s^j {\mathbf{w}_{\mathbf{h_{q_i}}}}^j, \mathbf{q}_{h_i}^j) < 0 \\ \mathbf{q}_{h_i}^j + s^j {\mathbf{w}_{\mathbf{h_{q_i}}}}^j, & \text{otherwise} \end{cases}$$

update $u_h^{j+1} = u_h^j + s^j w_{h_u}^j$ and set $j \leftarrow j + 1$

end

In the following experiments we use the parameters $\rho = 1$ for the scalar penalization of the augmented Lagrangian functional (4.4). In addition, the step size for the multiplier's update in ALG1 is fixed as $t^k = 1$ for all k-th iteration.

We compare the two Augmented Lagrangian algorithms, ALG2 and the modified version of ALG1 by solving the Bingham fluid problem in a pipe. The modified ALG1, enhanced with second-order information and a predictive strategy for identifying rigid zones in the fluid, accelerates the convergence process. This improvement is evident in the reduction on the number of iterations required to achieve a comparable level of the cost of the functional $\tilde{J}(u_h^k)$. Moreover, Table 4.1 shows two values of the parameter g. For $g = 0.2$ and the error in the constraint, $\|\nabla_h(u_h) - \mathbf{q}_h\|$, the modified ALG1 achieves a slightly smaller constraint error (6.79×10^{-4}) compared to ALG2 (9.33×10^{-4}) while requiring fewer iterations (16 versus 18) and producing identical plug flow

velocities (0.1169). Similarly, for $g = 0.4$, the modified ALG1 demonstrates competitive performance, with a constraint error of 9.44×10^{-4} compared to 9.90×10^{-4} for ALG2, and fewer iterations (43 versus 47). These results indicate that the modified ALG1 maintains comparable accuracy to ALG2 while generally requiring fewer iterations.

Second-order information in ALG1 provides an accurate descent direction, while the rigid-zone prediction strategy focus on key regions of the domain. As illustrated in Figure 4.2, the functional $\tilde{J}(u_h^k)$ decreases more rapidly with the modified ALG1 than with ALG2, highlighting the efficiency of combining second-order information and predictive strategies in tackling nonsmooth optimization problems in viscoplastic fluid dynamics.

Numerical performance of modified ALG1 vs ALG2					
$\frac{g}{\sqrt{2}}$	Aug.Lag- Alg.	$\|\nabla_h(u_k) - \mathbf{q}_h\|$	It.	Cost	plug flow velocity
0.2	ALG1 (mod)	6.79e-04	16	-0.077	0.1169
	ALG2	9.33e-04	18	-0.077	0.1169
0.4	ALG1 (mod)	9.44e-04	43	-0.00377	0.0192
	ALG2	9.90e-04	47	-0.00188	0.0192

Table 4.1: Performance of the Augmented Lagrangian Algorithms for different values of $\frac{g}{\sqrt{2}}$.

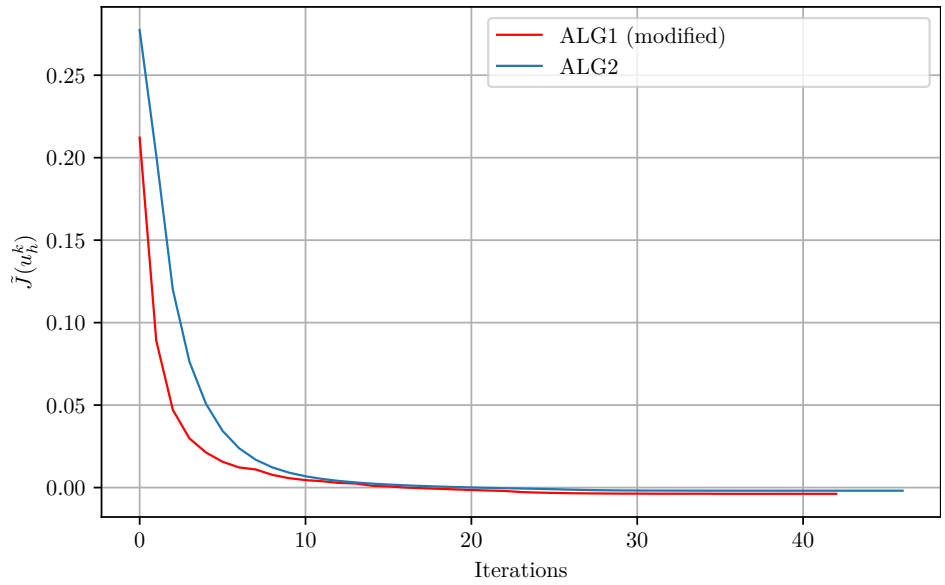


Figure 4.2: Comparison between the modified ALG1 and ALG2 for $\frac{g}{\sqrt{2}} = 0.4$.

Although the results of this first part are interesting from the theoretical point of view, in the next section, we will broaden the scope of the active-set strategy developed

in Section 4.2.5 to address a more extensive class of optimization problems involving the $\|\cdot\|_{1,2}$ norm.

4.3 General Group-sparse Problem: Formulation and Optimality Conditions

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a C^2-class function, not necessarily convex, with Lispchitz continuous gradient. We define the sparsity-promoting regularizer term as $\|\mathbf{u}\|_{1,2} = \sum_{i=1}^{p} \|\mathbf{u}_i\|_2$. Given a sparsity parameter $\sigma > 0$, we are interested in the following optimization problem:

$$\min_{\mathbf{u}\in\mathbb{R}^m} \psi(\mathbf{u}) := f(\mathbf{u}) + \sigma\|\mathbf{u}\|_{1,2}. \tag{GS}$$

We assume that the vector $\mathbf{u}^\top = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_i^\top, \ldots, \mathbf{u}_p^\top)$ is comprised of p subvectors $\mathbf{u}_i$, where each subvector belongs to $\mathbb{R}^{n_i}$, with $n_i \in \mathbb{N}$, for $i = 1, \ldots, p$, and $\sum_{i=1}^{p} n_i = m$. The subvector $\mathbf{u}_i$ is referred to as the i-th group of $\mathbf{u}$. In this section, the Euclidean inner product is denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|_2$ denotes the Euclidean norm.

To illustrate problem (GS), the most common example of applications that fit this formulation is the group LASSO problem:

$$\min_{\mathbf{u}\in\mathbb{R}^m} \psi(\mathbf{u}) := \frac{1}{2}\|A\mathbf{u} - \mathbf{b}\|_2^2 + \sigma\|\mathbf{u}\|_{1,2}.$$

This optimization problem is intended to solve an undetermined linear system with the $\ell_{1,2}$ regularization. It was introduced in [123] to investigate the selection of grouped variables in statistics. In this case, problem (GS) is given by $f(\mathbf{u}) := \frac{1}{2}\|A\mathbf{u} - b\|_2^2$, with $A \in \mathbb{R}^{l\times m}$. The group LASSO optimization problem extends the traditional Lasso problem by incorporating the $\|\cdot\|_{1,2}$ norm, which promotes sparsity across predefined groups of variables rather than individual coefficients. This approach is particularly useful in scenarios where the variables naturally form groups, such as in multi-task learning or hierarchical models. The objective is to minimize the least-squares loss function and to identify relevant groups while excluding irrelevant ones.

In this work, we assume f not necessarily convex. Therefore, we explore applications of problem (GS) in PDE-constrained optimization and nonlinear regression models.

4.3.1 Optimality condition

Function ψ, from problem (GS), is directionally differentiable thanks to the twice continuous differentiability of f and the directional differentiability of the norm $\|\cdot\|_{1,2}$. Let us recall that the directional derivative of ψ at $\mathbf{u}$ in a direction $\mathbf{v}$ is denoted by $\psi'(\mathbf{u}, \mathbf{v})$.

Furthermore, the continuous differentiability of f ensures that the Fréchet derivative coincides with the directional derivative, i.e., $f'(\mathbf{u})(\mathbf{v}) = f'(\mathbf{u}, \mathbf{y})$. Additionally, due to the convexity of $\|\cdot\|_{1,2}$ and by applying Theorem 2.12, we conclude that the Clarke's generalized directional derivative coincides with the directional derivative:

$$\psi^{'}(\mathbf{u}, \mathbf{v}) = \psi^{\circ}(\mathbf{u}, \mathbf{v}). \tag{4.28}$$

Let us review the necessary optimality conditions for problem (**GS**).

Utilizing the convexity of the $\|\cdot\|_{1,2}$ norm, from Theorem 2.12 we get that $\partial_C\|\cdot\|_{1,2} = \partial\|\cdot\|_{1,2}$. In addition, since f is a C^2-class function, by applying Theorems 2.11 and 2.13 we conclude that for a local minimum of ψ, $\mathbf{u}^* \in \mathbb{R}^m$, then

$$-\nabla f(\mathbf{u}^*) \in \sigma\partial\|\cdot\|_{1,2}(\mathbf{u}^*), \tag{4.29}$$

which means that the necessary optimality condition reduces to the following variational inequality:

$$\langle \nabla f'(\mathbf{u}^*), \mathbf{y} - \mathbf{u}^* \rangle + \sigma\|\mathbf{y}\|_{1,2} - \sigma\|\mathbf{u}^*\|_{1,2} \geq 0, \text{ for all } \mathbf{y} \in \mathbb{R}^m. \tag{4.30}$$

Moreover, thanks to the convexity of the $\|\cdot\|_{1,2}$ norm, from Proposition 2.1 we deduce that (4.30) is equivalent to

$$\psi'(\mathbf{u}^*, \mathbf{y} - \mathbf{u}^*) \geq 0, \text{ for all } \mathbf{y} \in \mathbb{R}^m.$$

Additionally, the subdifferential of the sparsity norm is characterized by $\partial\|\cdot\|_{1,2}(\mathbf{u}) = \{\mathbf{g} \in \mathbb{R}^m : \mathbf{g}_i \in \partial\|\cdot\|_2(\mathbf{u}_i), \forall i = 1, \cdots, p\}$. Here, the i-th group $\mathbf{g}_i \in \mathbb{R}^{n_i}$ satisfies:

$$\mathbf{g}_i \in \partial\|\cdot\|_2(\mathbf{u}_i) \Leftrightarrow \begin{cases} \|\mathbf{g}_i\|_2 \leq 1, & \text{if } \|\mathbf{u}_i\|_2 = 0, \\ \mathbf{g}_i = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_2}, & \text{if } \|\mathbf{u}_i\|_2 \neq 0. \end{cases} \tag{4.31}$$

Henceforth, we will drop the subscript in the Euclidean norm.

Thanks to (4.31), we rewrite the necessary optimality condition (4.29) as follows:

$$\|\nabla_i f(\mathbf{u}^*)\| \leq \sigma, \qquad \text{if } \|\mathbf{u}_i^*\| = 0, \tag{4.32a}$$

$$-\nabla_i f(\mathbf{u}^*) = \sigma\frac{\mathbf{u}_i^*}{\|\mathbf{u}_i^*\|}, \qquad \text{if } \|\mathbf{u}_i^*\| \neq 0 \tag{4.32b}$$

where $\nabla_i f(\mathbf{u}^*) := \left(\frac{\partial f}{\partial \mathbf{u}_{i_1}}(\mathbf{u}^*), \cdots, \frac{\partial f}{\partial \mathbf{u}_{i_{n_i}}}(\mathbf{u}^*)\right)^\top$ corresponds to the components belonging to the i-th group of the gradient $\nabla f(\mathbf{u}^k)$.

Next, from Definition 2.5, we have the following characterization.

Definition 4.1. *A point is said to be stationary if and only if the necessary optimality condition* (4.32) *is verified.*

4.3.2 Application Examples in Nonconvex optimization

We motivate problem (**GS**) by reviewing some important application examples that arise in nonlinear least squares regression and PDE-constrained optimization.

Nonlinear least-squares with group sparsity

Nonlinear least-squares problems with group sparsity are crucial in various fields due to their ability to model complex, real-world phenomena while enforcing structured sparsity in solutions. By incorporating group sparsity, these models not only improve interpretability by identifying relevant groups of variables but also enhance computational efficiency by reducing the dimensionality of the problem. Furthermore, the nonlinearity allows for more accurate modeling of intricate relationships within the data, leading to better predictive performance and more robust solutions in scenarios where traditional linear methods fall short.

An example of a nonlinear least-squares problem is formulated by the following model:

$$\min_{\mathbf{u}\in\mathbb{R}^m} \psi(\mathbf{u}) := \frac{1}{2}\|R(\mathbf{u})\|_2^2 + \sigma\|\mathbf{u}\|_{1,2}. \tag{NLS}$$

Here, $f(\mathbf{u}) = \frac{1}{2}\|R(\mathbf{u})\|^2$, with a differentiable function $R : \mathbb{R}^m \to \mathbb{R}^l$. For $j = 1, \cdots, l$, each $R_j(\mathbf{u}) \in \mathbb{R}$ corresponds, for instance, to a sigmoid function of the form $R_j(\mathbf{u}) = b_j - \dfrac{1}{1+e^{-\mathbf{a}_j^\top \mathbf{u}}}$, where $b_j \in \{0, 1\}$ and $\mathbf{a}_j \in \mathbb{R}^m$. This model arises in the context of logistic regression where $\mathbf{a}_j \in \mathbb{R}^m$ represents a feature vector, b_j is the corresponding label, and $\mathbf{u} \in \mathbb{R}^m$ is the parameter vector to be estimated. This formulation models the prediction error of the logistic function, i.e., it aims to minimize the squared residual norm where each component of the residual function $R_j(\mathbf{u})$ measures the difference between the observed label and the model prediction given by the logistic function. This formulation is widely used in classification tasks as in [12].

Problem (NLS) has the characteristics we aim to address in problem (**GS**). Specifically, it serves as an example of a composite optimization problem that includes both a differentiable but non-convex term and a non-differentiable but convex term. Additionally, the group sparsity term $\|\mathbf{u}\|_{1,2}$ promotes sparsity on the subvectors $\mathbf{u}_i$, i.e., if sparsity occurs at the i-th group, then $\mathbf{u}_i = \mathbf{0}$ (see [9] and [123]).

Nonconvex support vector machines

Support Vector Machines (SVMs) are supervised learning models used for classification. The core idea is to find a hyperplane that best separates the two classes in a high-dimensional space, maximizing the margin between them. Given training data and labels, the SVM optimization problem seeks a vector such that the decision function correctly classifies the samples while maintaining a maximal margin.

Consider the following support vector machine problem:

$$\min_{\mathbf{u}\in\mathbb{R}^m} \psi(\mathbf{u}) := \frac{1}{l}\sum_{j=1}^{l}\Big(1-\tanh(b_j\langle \mathbf{a}_j,\mathbf{u}\rangle)\Big)+\sigma\|\mathbf{u}\|_{1,2}. \tag{SVM}$$

Here, $b_j \in \{-1,1\}$ are labels and $\mathbf{a}_j \in \mathbb{R}^m$ are data points containing the information of the feature vectors for $j = 1,\cdots,l$. The term tanh is the hyperbolic tangent function, widely used in machine learning to model the nonlinearity of the data. The objective function in (SVM) is nonconvex due to the presence of the tanh function.

This type of model has been applied in binary classification tasks. We refer, for instance, to [24] for diabetes prediction. In particular, the diabetes database will be used to illustrate the performance of the proposed algorithm in that kind of problems.

Elliptic PDE constrained optimization

In PDE-constrained optimization, the variables belong to a function-space, for instance the space of square integrable functions $L^2(\Omega)$. Following the discretize-then-optimize paradigm, we transition from a continuous to a discrete framework by approximating the function spaces by some finite-dimensional ones. This approximation is achieved through suitable techniques (e.g., the finite element method). After a discretization procedure, the resulting approximating problem can be formulated in the form of (**GS**).

Let $\Omega \subset \mathbb{R}^N$ $(N > 1)$ be an open bounded domain with Lipschitz boundary Γ, and consider the following semilinear elliptic PDE:

$$\begin{aligned} -\Delta y + a(y) &= u \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.33}$$

where the control u is of distributed kind. Assuming that the state y is the unique solution to the PDE, we may define the control-to-state mapping by $S : L^2(\Omega) \to L^2(\Omega)$, where $u \mapsto S(u) = y$.

Using the control-to-state mapping and considering a quadratic cost functional, a

reduced PDE optimal control problem is given by

$$\min_{u\in L^2(\Omega)} \psi := \frac{1}{2}\|S(u) - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} + \sigma\|u\|_{1,2}, \tag{OCP}$$

where $y_d \in L^2(\Omega)$ stands for a given desired state and α is a positive real constant. The norm $\|u\|_{1,2}$ promotes controls with certain structured sparsity patterns. These patterns, known as "directionally sparse controls" consider the problem's structure, activating only certain directions in the control space to achieve the optimal solution. In other words, the optimal control strategy involves applying controls along specific directions while keeping other directions sparse. This approach was introduced in [64], where optimal control problems involving linear elliptic and parabolic equations were studied. An extension of this research to the semilinear case was done in [23].

To illustrate the concept of directional sparsity, we utilize the following partitioning of Ω as proposed in [64, Sec. 1]: For some $1 \leq \bar{m} < N$, $\Omega = \Omega_1 \times \Omega_2 \subset \mathbb{R}^{\bar{m}} \times \mathbb{R}^{N-\bar{m}}$, with

$$\begin{aligned} \Omega_1 &= \{x_1 \in \mathbb{R}^{\bar{m}} : \exists x_2 \in \mathbb{R}^{N-\bar{m}} \text{ with } (x_1, x_2) \in \Omega\}, \\ \Omega_2(x_1) &= \{x_2 \in \mathbb{R}^{N-\bar{m}} : (x_1, x_2) \in \Omega\} \text{ for } x_1 \in \Omega_1. \end{aligned}$$

In two dimensions ($N = 2$, $\bar{m} = 1$), we may consider, for instance, $\Omega_1 = \Omega_2 = (0, 1)$, i.e., $\Omega = (0, 1) \times (0, 1)$. Thus, $\{x_1\} \times \Omega_2(x_1)$ represents the vertical cross section of Ω at $x_1 \in \Omega_1$. In this setting, the directional sparsity norm is given by

$$\|u\|_{1,2} := \int_{\Omega_1} \left(\int_{\Omega_2(x_1)} u^2(x_1, x_2)\, dx_2 \right)^{1/2} dx_1 = \int_0^1 \left(\int_0^1 u^2(x_1, x_2)\, dx_2 \right)^{1/2} dx_1.$$

The sparsity patterns promoted by this norm are given over the vertical cross sections of the domain, see Figure 4.3.

Time-dependent PDE constrained optimization

In this case, we consider a domain given by the space-time cylinder $Q = \Omega \times (0, T)$ with $\Omega = (0, 1) \times (0, 1)$. Here, $N = 3$ and $\bar{m} = 2$. The chosen control space is $L^2(Q)$ and the control-to-state mapping S assigns to each control u the solution $y \in W(0, T) \cap C(\bar{Q})$ to the following parabolic semilinear PDE:

$$\begin{aligned} y_t - \nabla \cdot (\kappa \nabla) y + a(y) &= u && \text{in } Q, \\ y &= 0 && \text{on } \Gamma \times (0, T), \\ y(\cdot, 0) &= 0 && \text{in } \Omega, \end{aligned}$$

where κ and the semilinear term $a(\cdot)$ satisfy suitable conditions (see [113, Chapter 5]).

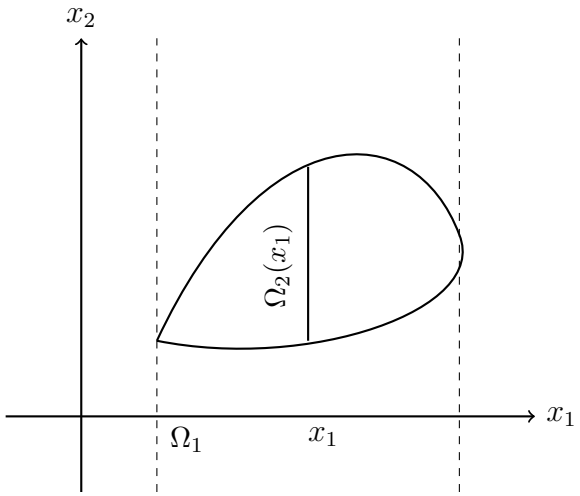


Figure 4.3: Directional sparsity

With help of the control-to-state mapping and considering again a tracking type cost functional, we arrive at the following optimal control problem:

$$\min_{u \in L^2(Q)} \psi := \frac{1}{2}\|S(u) - y_d\|^2_{L^2(Q)} + \frac{\alpha}{2}\|u\|^2_{L^2(Q)} + \sigma\|u\|_{1,2}, \tag{OCP}$$

where $y_d \in L^2(Q)$ is the desired state and the sparsity norm reads:

$$\|u\|_{1,2} := \int_\Omega \left(\int_0^T u^2(x,t)\, dt \right)^{1/2} dx.$$

This is called directional sparsity since the sparsity structure is induced in the space domain and fixed over time, meaning that controls are sparse in space and the sparsity patterns does not change in time. Alternatively, sparsity can be considered in the time variable.

Following the discretize-then-optimize approach, the transition from the continuous to the discrete problem requires space-time discretization.

Note that, in the presence of semi-linear partial differential equations, the reduced optimal control problem (OCP) is not convex [113].

The results presented in the following section were first introduced and proved in [40].

4.3.3 The Group Sparse Descent Method

In this section, we present the constitutive components of the proposed *Group Sparse Descent Method (GSDM)* introduced in [40]. Summarizing the main steps, we begin

by computing the steepest descent direction of the problem (**GS**). Next, the proposed active-set prediction phase is introduced, and its main theoretical implications are explained. Based on the active-set prediction strategy, a modified descent direction is then introduced, which is subsequently adjusted using generalized second-order information of the cost function in (**GS**).

Steepest descent direction

Recalling the directional differentiability of ψ, in this section we will determine the steepest descent direction of ψ at $\mathbf{u}$.

From Theorem 2.14 and given the relation in (4.28) we have the following definition.

Definition 4.2. *A vector $\mathbf{z} \in \mathbb{R}^m$ is a descent direction for ψ at $\mathbf{u}$ if $\mathbf{z}$ satisfies the condition:*

$$\psi^{'}(\mathbf{u}, \mathbf{z}) < 0.$$

Remark 4.1. *From Corollary 2.2 we have that if $\mathbf{u}$ is a nonstationary point, then there exists a descent direction $\mathbf{z}$ for ψ at $\mathbf{u}$. Therefore, either the necessary optimality condition (4.32) holds or there exists a descent direction $\mathbf{z} \in \mathbb{R}^m$ for ψ at $\mathbf{u}$.*

Moreover, as discussed in Section 2.2.2 (subproblem (2.17)), due to (4.28) the steepest descent direction of ψ at $\mathbf{u}$ is characterized as the unique minimizer of $\psi^{'}(\mathbf{u}, \cdot)$ over the ball $B(0, 1)$ in $\mathbb{R}^m$, i.e., the steepest descent direction $\hat{\mathbf{z}}$ can be computed by solving:

$$\hat{\mathbf{z}} = \underset{\|\mathbf{z}\| \leq 1}{\arg\min}\, \psi'(\mathbf{u}, \mathbf{z})$$

in the Euclidean norm $\|\cdot\|$. In our case, the groupwise structure of the $\|\cdot\|_{1,2}$ norm penalizer motivate us to replace $\|\mathbf{z}\|$ with the dual norm $\|\mathbf{z}\|_{\infty,2}$. Thus the steepest descent direction is considered in the following sense:

$$\hat{\mathbf{z}} = \underset{\|\mathbf{z}\|_{\infty,2} \leq 1}{\arg\min}\, \psi'(\mathbf{u}, \mathbf{z}). \tag{4.34}$$

Let us denote the Euclidean norm of each i-th group by $h(\mathbf{u}_i) := \|\mathbf{u}_i\|$. Then, we have that $\|\mathbf{u}\|_{1,2} = \sum_{i=1}^{p} h(\mathbf{u}_i)$. Moreover, the directional derivative of h at $\mathbf{u}_i$ in the direction $\mathbf{z}_i$, is given by:

$$h'(\mathbf{u}_i, \mathbf{z}_i) = \begin{cases} \dfrac{\langle \mathbf{u}_i, \mathbf{z}_i \rangle}{\|\mathbf{u}_i\|}, & \text{if } \mathbf{u}_i \neq \mathbf{0}, \\ \|\mathbf{z}_i\|, & \text{otherwise.} \end{cases} \tag{4.35}$$

Consequently, (4.34) becomes

$$\hat{\mathbf{z}} = \underset{\|\mathbf{z}\|_{\infty,2}\leq 1}{\arg\min} \left\{ \langle \nabla f(\mathbf{u}), \mathbf{z} \rangle + \sigma \sum_{i=1}^{p} h'(\mathbf{u}_i, \mathbf{z}_i) \right\}.$$

Thanks to the separable structure of the previous problem, we derive the following subproblems:

$$\bar{\mathbf{z}}_i = \underset{\|\mathbf{z}_i\|\leq 1}{\arg\min} \left\{ \langle \nabla_i f(\mathbf{u}), \mathbf{z}_i \rangle + \sigma h'(\mathbf{u}_i, \mathbf{z}_i) \right\}, \ \forall i = 1, \ldots, p. \tag{4.36}$$

We shall determine the steepest descent direction, $\bar{\mathbf{z}}$, group-wise, by solving (4.36) for each group indexed by $i \in \{1, \cdots, p\}$.

In addition, if $\bar{\mathbf{z}}_i$ is the solution of (4.36), for $i = 1, \ldots, p$, each $\bar{\mathbf{z}}_i$ can be expressed as

$$\bar{\mathbf{z}}_i = -\frac{\bar{\mathbf{v}}_i}{\|\bar{\mathbf{v}}_i\|},$$

where $\bar{\mathbf{v}}_i \in \nabla_i f(\mathbf{u}) + \sigma \partial \|\cdot\|(\mathbf{u}_i)$ is the minimum norm subgradient (see Section 2.2.2 - equation (2.18)).

Next, in order to describe the solution to (4.36), we introduce the following index-sets that characterize $\bar{\mathbf{z}}_i = -\frac{\bar{\mathbf{v}}_i}{\|\bar{\mathbf{v}}_i\|}$ on the *active* and *inactive* groups:

$$\begin{aligned}
\mathcal{A}_0(\mathbf{u}) &:= \{i : \|\mathbf{u}_i\| = 0 \text{ and } \|\nabla_i f(\mathbf{u})\| \leq \sigma\}, \\
\mathcal{A}_\nabla(\mathbf{u}) &:= \{i : \|\mathbf{u}_i\| = 0 \text{ and } \|\nabla_i f(\mathbf{u})\| > \sigma\}, \\
\mathcal{I}(\mathbf{u}) &:= \{i : \|\mathbf{u}_i\| \neq 0\}.
\end{aligned}$$

A group indexed in the set $\mathcal{A}_0(\mathbf{u})$ satisfies the necessary optimality condition (4.32a). Henceforth, we will refer to this index-set as the active index-set. Accordingly, $\mathcal{I}(\mathbf{u})$ corresponds to the inactive index-set.

Theorem 4.3. *The unique solution to* (4.36) *is given by* $\bar{\mathbf{z}}$ *with components* $\bar{\mathbf{z}}_i$ *that satisfy*

$$\bar{\mathbf{z}}_i = \begin{cases} \dfrac{-\nabla_i f(\mathbf{u}) - \sigma \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|}}{\left\| \nabla_i f(\mathbf{u}) + \sigma \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} \right\|}, & \text{if } i \in \mathcal{I}(\mathbf{u}), \\ \mathbf{0}, & \text{if } i \in \mathcal{A}_0(\mathbf{u}), \\ \dfrac{-\nabla_i f(\mathbf{u})}{\left\| \nabla_i f(\mathbf{u}) \right\|}, & \text{if } i \in \mathcal{A}_\nabla(\mathbf{u}). \end{cases} \tag{4.37}$$

for all $i = 1, \cdots, p$.

Proof. Let us start with the inactive index set $\mathcal{I}(\mathbf{u})$. From (4.35) and since $\|\cdot\|$ is differentiable at any $\mathbf{u}_i \neq 0$, it is clear that the solution to (4.36) is attained at

$$\bar{\mathbf{z}}_i = -\frac{\left\langle \nabla_i f(\mathbf{u}) + \sigma \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} \right\rangle}{\left\| \nabla_i f(\mathbf{u}) + \sigma \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} \right\|}.$$

On the other hand, if $\mathbf{u}_i = \mathbf{0}$, using (4.35), problem (4.36) reads as

$$\bar{\mathbf{z}}_i = \underset{\|\mathbf{z}_i\| \leq 1}{\arg\min} \left\{ \langle \nabla_i f(\mathbf{u}), \boldsymbol{\nu}_i \rangle + \sigma \|\mathbf{z}_i\| \right\}.$$

Here, two possibilities arise:

- If $i \in \mathcal{A}_0(\mathbf{u})$, then it follows that

  $$\langle \nabla_i f(\mathbf{u}), \mathbf{z}_i \rangle + \sigma \|\mathbf{z}_i\| \geq -\|\nabla_i f(\mathbf{u})\| \|\mathbf{z}_i\| + \sigma \|\mathbf{z}_i\| = (\sigma - \|\nabla_i f(\mathbf{u})\|) \|\mathbf{z}_i\| \geq 0,$$

  for any $\mathbf{z}_i$ with $\|\mathbf{z}_i\| \leq 1$. Consequently, the minimizer of (4.36) is given by $\bar{\mathbf{z}}_i = \mathbf{0}$, i.e., the descent direction vanishes.

- If $i \in \mathcal{A}_\nabla(\mathbf{u})$, then

  $$\langle \nabla_i f(\mathbf{u}), \mathbf{z}_i \rangle + \sigma \|\mathbf{z}_i\| \geq -\|\nabla_i f(\mathbf{u})\| \|\mathbf{z}_i\| + \sigma \|\mathbf{z}_i\| = -\|\nabla_i f(\mathbf{u})\| + \sigma,$$

  for any $\mathbf{z}_i$ with $\|\mathbf{z}_i\| \leq 1$. Therefore, the minimizer of (4.36) is given by $\bar{\mathbf{z}}_i = -\frac{\nabla_i f(\mathbf{u})}{\|\nabla_i f(\mathbf{u})\|}$, since with that choice the lower bound is attained.

Altogether, we arrive at the steepest descent direction given in (4.37). □

Remark 4.2. *For all $i = 1, \cdots, p$, the minimum norm subgradient $\bar{\mathbf{v}}_i$ related to the cost function (4.36) is of the form $\bar{\mathbf{v}}_i = \nabla_i f(\mathbf{u}) + \sigma \bar{\boldsymbol{\xi}}_i$, with $\bar{\boldsymbol{\xi}}_i \in \partial \|\cdot\|(\mathbf{u}_i)$, such that*

$$\bar{\mathbf{v}}_i = \begin{cases} \nabla_i f(\mathbf{u}) + \sigma \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|}, & \text{if } i \in \mathcal{I}(\mathbf{u}), \\ \mathbf{0}, & \text{if } i \in \mathcal{A}_0(\mathbf{u}), \\ \nabla_i f(\mathbf{u}) - \sigma \frac{\nabla_i f(\mathbf{u})}{\|\nabla_i f(\mathbf{u})\|}, & \text{if } i \in \mathcal{A}_\nabla(\mathbf{u}). \end{cases} \tag{4.38}$$

Consequently, the steepest descent direction coincides with the negative minimum norm subgradient direction.

Algorithms for minimizing nondifferentiable functionals based on the computation of the directional derivative to achieve the steepest descent can be traced back to [15] and [88]. However, steepest descent methods exhibit poor performance for nonconvex functionals [3], especially when combined with sparsity terms.

Active-set prediction phase

Identifying the active set during the computation of solutions for sparse-optimization problems has a significant impact on the performance of optimization algorithms (see, e.g., [21, 39, 87, 105]). The purpose of active-set strategies is to iteratively identify and update the set of groups that become active or inactive at each iteration. This procedure reduces the optimization complexity by limiting the number of variables under consideration and enhances the efficiency of the optimization algorithm by accelerating convergence in the inactive index set.

In the first part of this chapter, we introduced an active-set strategy for identifying the active components of a convex problem, such as the Bingham flow in a pipe. This approach was based on analyzing the angle between two consecutive iterations. In this section, we extend this concept to problem (**GS**) by examining its optimality condition. We argue that, as the solution is approached to a local optimum, the angle between consecutive iterations remains positive. Therefore, the angle between the current iterate and the descent direction for each group remains positive as well. This result was first proved in [40]. Building on this result, the active-set phase aims to iteratively identify the groups that become active or inactive during each iteration.

Consequently, the proposed prediction strategy is motivated by the following observation: from the necessary optimality condition (4.32), we obtain, for the i-th inactive group ($\mathbf{u}_i^* \neq \mathbf{0}$), that

$$\left\langle -\frac{1}{\sigma}\nabla_i f(\mathbf{u}^*), \mathbf{u}_i^* \right\rangle = \|\mathbf{u}_i^*\| = \left\langle \frac{\mathbf{u}_i^*}{\|\mathbf{u}_i^*\|}, \mathbf{u}_i^* \right\rangle,$$

or, equivalently,

$$\left\langle -\nabla_i f(\mathbf{u}^*) - \sigma\frac{\mathbf{u}_i^*}{\|\mathbf{u}_i^*\|}, \mathbf{u}_i^* \right\rangle = 0. \tag{4.39}$$

The main idea of the proposed active-set strategy consists in using (4.39), in an iterative manner, to identify active groups.

Specifically, the left hand side argument of the inner product in (4.39) corresponds to the steepest descent direction (see (4.38)), which vanishes for inactive indices. Consequently, if at a given iterate $\mathbf{u}_i^k \neq 0$, the steepest descent direction is non-zero, considering the update $\mathbf{u}_i^{k+1} = \mathbf{u}_i^k - \bar{\mathbf{v}}_i^k$, the following holds:

$$\begin{aligned} \left\langle \mathbf{u}_i^{k+1}, \mathbf{u}_i^k \right\rangle = \left\langle \mathbf{u}_i^k - \bar{\mathbf{v}}_i^k, \mathbf{u}_i^k \right\rangle &= \left\langle \mathbf{u}_i^k - \left(\nabla_i f(\mathbf{u}^k) + \sigma\tfrac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}\right), \mathbf{u}_i^k \right\rangle \\ &= \|\mathbf{u}_i^k\|^2 + \left\langle -\nabla_i f(\mathbf{u}^k) - \sigma\tfrac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}, \mathbf{u}_i^k \right\rangle. \end{aligned} \tag{4.40}$$

If the iterates are close to the optimum and $i \in \mathcal{I}(\mathbf{u}^*)$, the term $\left\langle -\nabla_i f(\mathbf{u}^k) - \sigma\frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}, \mathbf{u}_i^k \right\rangle$

approaches zero, and ultimately vanishes at the optimum (see (4.39)), while $\|\mathbf{u}_i^k\|$ remains strictly positive. This implies that the inner product $\left\langle \mathbf{u}_i^{k+1}, \mathbf{u}_i^k \right\rangle$ should remain positive for inactive groups. As a consequence, the active-set prediction phase is based on the violation of this positivity.

Thus, the active-set prediction phase is based on the following index-set:

$$\mathcal{J}^k = \{i : \langle \mathbf{u}_i^k, -\bar{\mathbf{v}}_i^k \rangle < 0\}, \tag{4.41}$$

which contains the indices of the groups that are expected to become active at the next iteration. The rationale behind this definition is that these are the indices where the condition $\left\langle \mathbf{u}_i^{k+1}, \mathbf{u}_i^k \right\rangle > 0$ may be violated, depending on the value of $\|\mathbf{u}_i^k\|$. This strategy also coincides, in the specific case of sparse-optimization problems with the ℓ_1-norm, with the orthantwise strategy developed in [1, 21, 39].

We define the *predictive active* index-set, consequently, as

$$\tilde{\mathcal{A}}^k := \mathcal{A}_0(\mathbf{u}^k) \cup \mathcal{A}_\nabla(\mathbf{u}^k) \cup \mathcal{J}^k. \tag{4.42}$$

Moreover the *predictive inactive* index-set is given by

$$\tilde{\mathcal{I}}^k := \mathcal{I}(\mathbf{u}^k) \setminus \mathcal{J}^k. \tag{4.43}$$

Hereafter, we will omit the argument dependence in the notation of the index-sets. For example, we write $\mathcal{A}_0^k$ instead of $\mathcal{A}_0(\mathbf{u}^k)$.

These sets enable us to introduce a modified direction, denoted by $-\tilde{\mathbf{d}}_i^k$, which induces sparsity in $\mathbf{u}^{k+1}$. This direction is defined as:

$$-\tilde{\mathbf{d}}_i^k = \begin{cases} -\mathbf{u}_i^k, & \text{if } i \in \mathcal{J}^k \\ -\bar{\mathbf{v}}_i^k, & \text{otherwise} \end{cases} = \begin{cases} -\mathbf{u}_i^k, & \text{if } i \in \mathcal{J}^k, \\ -\nabla_i f(\mathbf{u}^k) - \sigma \dfrac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}, & \text{if } i \in \tilde{\mathcal{I}}^k = \mathcal{I}^k \setminus \mathcal{J}^k, \\ -\nabla_i f(\mathbf{u}^k) + \sigma \dfrac{\nabla_i f(\mathbf{u}^k)}{\|\nabla_i f(\mathbf{u}^k)\|}, & \text{if } i \in \mathcal{A}_\nabla^k, \\ 0, & \text{if } i \in \mathcal{A}_0^k. \end{cases} \tag{4.44}$$

Remark 4.3. *In the update step $\mathbf{u}_i^{k+1} = \mathbf{u}_i^k - \tilde{\mathbf{d}}_i^k$, the modified direction will set the groups in $\mathcal{J}^k$ to exactly zero as follows:*

$$\mathbf{u}_i^{k+1} = \mathbf{u}_i^k - \tilde{\mathbf{d}}_i^k = \begin{cases} \mathbf{0}, & \textit{if } i \in \mathcal{J}^k \\ \mathbf{u}_i^k - \bar{\mathbf{v}}_i^k, & \textit{otherwise.} \end{cases} \tag{4.45}$$

If, at a given iteration, one or more indices are added to the predictive active set

$\tilde{\mathcal{A}}^k$, the size of the inactive index-set $\mathcal{I}^k$ is reduced. Thus, we can compute second-order information only for those groups within the predicted inactive set $\tilde{\mathcal{I}}^k$, i.e., we incorporate second-order information in terms of a reduced second-order matrix. The larger the predictive $\tilde{\mathcal{A}}^k$ active set becomes, the smaller the second-order system required to compute the descent direction.

Ultimately, the updated solution (4.45) can be interpreted as a projection onto the origin, whenever that (4.40) holds.

The following result establishes that the modified direction $-\tilde{\mathbf{d}}_i^k$ is indeed a descent direction.

Proposition 4.1. *Let $\mathbf{u}^k$ be a non-stationary point. Then, the modified direction $-\tilde{\mathbf{d}}^k$, defined by (4.44), satisfies*

$$\psi'(\mathbf{u}^k, -\tilde{\mathbf{d}}^k) = -\sum_{i \in \mathcal{A}_\nabla^k \cup \tilde{\mathcal{I}}^k} \|\bar{\mathbf{v}}_i^k\|^2 < 0.$$

Proof. The directional derivative of ψ at $\mathbf{u}^k$, in the direction $-\tilde{\mathbf{d}}^k$, is given by

$$\psi'(\mathbf{u}^k, -\tilde{\mathbf{d}}^k) = \langle \nabla f(\mathbf{u}^k), -\tilde{\mathbf{d}}^k \rangle + \sigma \sum_{i=1}^{p} h'(\mathbf{u}_i^k, -\tilde{\mathbf{d}}_i^k). \tag{4.46}$$

We rewrite (4.46) by separating the set of group indices $\{i : 1 \leq i \leq p\}$ into $\mathcal{J}^k$ and its complement $\mathcal{J}^{k\complement}$. Then, (4.35) and (4.44) imply that:

$$\begin{aligned} \psi'(\mathbf{u}^k, -\tilde{\mathbf{d}}^k) \;=\; & \sum_{\substack{i \in \mathcal{J}^{k\complement} \\ \mathbf{u}_i^k = \mathbf{0}}} \langle \nabla_i f(\mathbf{u}^k), -\bar{\mathbf{v}}_i^k \rangle + \sigma \|\bar{\mathbf{v}}_i^k\| + \sum_{\substack{i \in \mathcal{J}^{k\complement} \\ \mathbf{u}_i^k \neq \mathbf{0}}} \langle \nabla_i f(\mathbf{u}^k), -\bar{\mathbf{v}}_i^k \rangle + \sigma \frac{\langle \mathbf{u}_i^k, -\bar{\mathbf{v}}_i^k \rangle}{\|\mathbf{u}_i^k\|} \\ & + \sum_{\substack{i \in \mathcal{J}^k \\ \mathbf{u}_i^k \neq \mathbf{0}}} \langle \nabla_i f(\mathbf{u}^k), -\mathbf{u}_i^k \rangle + \sigma \frac{\langle \mathbf{u}_i^k, -\mathbf{u}_i^k \rangle}{\|\mathbf{u}_i^k\|}. \end{aligned} \tag{4.47}$$

Recall that $i \in \mathcal{J}^k$ implies $\mathbf{u}_i^k \neq \mathbf{0}$. Hence, the sum over the set $\mathcal{J}^k$, when $\mathbf{u}_i^k = \mathbf{0}$, is not included in (4.47) because this set is empty. Moreover, since $\mathcal{A}_0^k \cup \mathcal{A}_\nabla^k \subset \mathcal{J}^{k\complement}$, we may rewrite the first term on the right-hand side of (4.47) as follows:

$$\sum_{\substack{i \in \mathcal{J}^{k\complement} \\ \mathbf{u}_i^k = \mathbf{0}}} \langle \nabla_i f(\mathbf{u}^k), -\bar{\mathbf{v}}_i^k \rangle + \sigma \|\bar{\mathbf{v}}_i^k\| = \sum_{i \in \mathcal{A}_0^k \cup \mathcal{A}_\nabla^k} \langle \nabla_i f(\mathbf{u}^k), -\bar{\mathbf{v}}_i^k \rangle + \sigma \|\bar{\mathbf{v}}_i^k\|.$$

Since $\bar{\mathbf{v}}_i^k = \mathbf{0}$, for all $i \in \mathcal{A}_0^k$, the directional derivative (4.47) becomes

$$\begin{aligned}\psi'(\mathbf{u}^k, -\tilde{\mathbf{d}}^k) &= \sum_{i\in\mathcal{A}_\nabla^k} \langle \nabla_i f(\mathbf{u}^k), -\bar{\mathbf{v}}_i^k\rangle + \sigma\|\bar{\mathbf{v}}_i^k\| + \sum_{\substack{i\notin\mathcal{J}^k\\ \mathbf{u}_i^k\neq\mathbf{0}}} \left\langle \nabla_i f(\mathbf{u}^k) + \sigma\frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}, -\bar{\mathbf{v}}_i^k\right\rangle \\ &\quad + \sum_{\substack{i\in\mathcal{J}^k\\ \mathbf{u}_i^k\neq\mathbf{0}}} \left\langle \nabla_i f(\mathbf{u}^k) + \sigma\frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}, -\mathbf{u}_i^k\right\rangle.\end{aligned} \tag{4.48}$$

Let us focus on the first sum over $\mathcal{A}_\nabla^k$ in (4.48). From the definition of $\bar{\mathbf{v}}_i^k$, given in (4.38), we get that:

$$\begin{aligned}\sum_{i\in\mathcal{A}_\nabla^k} \langle \nabla_i f(\mathbf{u}^k), -\bar{\mathbf{v}}_i^k\rangle + \sigma\|\bar{\mathbf{v}}_i^k\| &= \sum_{i\in\mathcal{A}_\nabla^k} \left\langle \nabla_i f(\mathbf{u}^k), -\left(1 - \frac{\sigma}{\|\nabla_i f(\mathbf{u}^k)\|}\right)\nabla_i f(\mathbf{u}^k)\right\rangle \\ &\quad + \sum_{i\in\mathcal{A}_\nabla^k} \sigma\left\|\left(1 - \frac{\sigma}{\|\nabla_i f(\mathbf{u}^k)\|}\right)\nabla_i f(\mathbf{u}^k)\right\|.\end{aligned} \tag{4.49}$$

Since $\|\nabla_i f(\mathbf{u}^k)\| > \sigma$, for all $i \in \mathcal{A}_\nabla^k$, then $\left(1 - \frac{\sigma}{\|\nabla_i f(\mathbf{u}^k)\|}\right) > 0$. Thus, we rewrite (4.49) as follows:

$$\begin{aligned}\sum_{i\in\mathcal{A}_\nabla^k} \langle \nabla_i f(\mathbf{u}^k), -\bar{\mathbf{v}}_i^k\rangle + \sigma\|\bar{\mathbf{v}}_i^k\| &= -\sum_{i\in\mathcal{A}_\nabla^k} \left(1 - \frac{\sigma}{\|\nabla_i f(\mathbf{u}^k)\|}\right)\|\nabla_i f(\mathbf{u}^k)\|^2 \\ &\quad + \sum_{i\in\mathcal{A}_\nabla^k} \sigma\left(1 - \frac{\sigma}{\|\nabla_i f(\mathbf{u}^k)\|}\right)\|\nabla_i f(\mathbf{u}^k)\| \\ &= -\sum_{i\in\mathcal{A}_\nabla^k} \|\nabla_i f(\mathbf{u}^k)\|^2 - 2\sigma\|\nabla_i f(\mathbf{u}^k)\| + \sigma^2 \\ &= -\sum_{i\in\mathcal{A}_\nabla^k} \left\|\nabla_i f(\mathbf{u}^k) - \sigma\frac{\nabla_i f(\mathbf{u}^k)}{\|\nabla_i f(\mathbf{u}^k)\|}\right\|^2 \\ &= -\sum_{i\in\mathcal{A}_\nabla^k} \|\bar{\mathbf{v}}_i^k\|^2.\end{aligned} \tag{4.50}$$

By replacing (4.50) and the definition of $\bar{\mathbf{v}}_i^k$ in the remaining terms of (4.48) we get that:

$$\psi'(\mathbf{u}^k, -\tilde{\mathbf{d}}^k) = -\sum_{i\in\mathcal{A}_\nabla^k} \|\bar{\mathbf{v}}_i^k\|^2 - \sum_{\substack{i\notin\mathcal{J}^k\\ \mathbf{u}_i^k\neq\mathbf{0}}} \|\bar{\mathbf{v}}_i^k\|^2 + \sum_{\substack{i\in\mathcal{J}^k\\ \mathbf{u}_i^k\neq\mathbf{0}}} \langle \bar{\mathbf{v}}_i^k, -\mathbf{u}_i^k\rangle,$$

where the last sum is also negative in view of (4.41). Therefore, from the definition of the index-set $\tilde{\mathcal{I}}^k = \mathcal{I}^k \setminus \mathcal{J}^k$ we conclude that

$$\psi'(\mathbf{u}^k, -\tilde{\mathbf{d}}^k) = -\sum_{i\in\mathcal{A}_\nabla^k} \|\bar{\mathbf{v}}_i^k\|^2 - \sum_{\substack{i\notin\mathcal{J}^k\\ \mathbf{u}_i^k\neq\mathbf{0}}} \|\bar{\mathbf{v}}_i^k\|^2 = -\sum_{i\in\mathcal{A}_\nabla^k\cup\tilde{\mathcal{I}}^k} \|\bar{\mathbf{v}}_i^k\|^2 < 0.$$

□

Incorporating second-order information

In the previous section, we described the phase that identifies active groups (sparse groups). The identification of the active index-set is the basis for constructing a reduced second-order matrix. If one or more indices are added to the active set at a given iteration, the size of the second-order matrix is reduced. In this section, we focus on improving the efficiency of the optimization algorithm by accelerating the search direction exclusively within the inactive index set.

Our strategy incorporates curvature information provided by generalized second-order derivatives from both terms within ψ, namely f and $\|\cdot\|_{1,2}$.

The second-order information for the smooth term f may be approximated locally by symmetric positive definite approximation of its Hessian (e.g. BFGS matrix), whereas for the nonsmooth part, the "nonsmooth curvature" is obtained indirectly, by using a regularization of the norm $\|\cdot\|_{1,2}$. This approach was proposed in [39] to get generalized second-order derivative for the ℓ_1 norm penalizer. Here, we extend this mechanism to the $\|\cdot\|_{1,2}$ norm.

Because the second-order system acts on the inactive groups only, the computational burden in each step of the algorithm is reduced proportionally to the number of variables that belong to sparse groups. We discuss the associated Newton-like system in detail in the next paragraphs.

Generalized second-order information for the $\|\cdot\|_{1,2}$ norm

Second-order information for the $\|\cdot\|_{1,2}$ norm can be approximated by smoothing. We introduce the local Huber regularization of the Euclidean norm $\|\cdot\|$ in $\mathbb{R}^{n_i}$, denoted by the function $h_\gamma : \mathbb{R}^{n_i} \to \mathbb{R}$ with parameter $\gamma > 0$, defined as:

$$h_\gamma(\mathbf{u}_i) = \begin{cases} \|\mathbf{u}_i\| - \frac{1}{2\gamma}, & if \ \|\mathbf{u}_i\| > \frac{1}{\gamma}, \\ \frac{\gamma}{2}\|\mathbf{u}_i\|^2, & if \ \|\mathbf{u}_i\| \leq \frac{1}{\gamma}, \end{cases}$$

for all $i = 1, \cdots, p$. This regularization is differentiable and its gradient is given by:

$$\nabla h_\gamma(\mathbf{u}_i) = \frac{\gamma \mathbf{u}_i}{\max(1, \gamma\|\mathbf{u}_i\|)}, \ \ \forall i = 1, \cdots, p.$$

The latter is not differentiable because of the max function. However, it is locally Lipschitz continuous and directionally differentiable. Thus, by Remark 2.2, it is Bouligand differentiable. In addition, since the max function is semismooth (see Example 2.2), the function $\nabla h_\gamma(\mathbf{u}_i)$ is semismooth by Proposition 2.4.

Using the rules of the Bouligand subdifferential [115], we compute a second-order

generalized derivative of h_γ. Let us denote by $\Gamma(\mathbf{u}_i)$ the element of the subdifferential given by

$$\Gamma(\mathbf{u}_i) = \begin{cases} \dfrac{1}{\|\mathbf{u}_i\|} I_{n_i} - \dfrac{\mathbf{u}_i \mathbf{u}_i^\top}{\|\mathbf{u}_i\|^3}, & \text{if } \|\mathbf{u}_i\| \geq \frac{1}{\gamma}, \\ \gamma I_{n_i}, & \text{if } \|\mathbf{u}_i\| < \frac{1}{\gamma}, \end{cases} \tag{4.51}$$

for all $i = 1, \ldots, p$. Here, I_{n_i} corresponds to the identity matrix in $\mathbb{R}^{n_i \times n_i}$. Furthermore, the second-order information of the regularized version of $\|\cdot\|_{1,2}$ is given by a symmetric positive semi-definite block-diagonal matrix $\Gamma(\mathbf{u})$, with p blocks of the form $\Gamma(\mathbf{u}_i)$ in the diagonal as follows:

$$\Gamma(\mathbf{u}) = \begin{pmatrix} \Gamma(\mathbf{u}_1) & & \\ & \ddots & \\ & & \Gamma(\mathbf{u}_p) \end{pmatrix}.$$

The positive semi-definiteness of $\Gamma(\mathbf{u})$ is proved in the following proposition.

Proposition 4.2. *The block-diagonal matrix $\Gamma(\mathbf{u}) \in \mathbb{R}^{m \times m}$, comprised of p blocks of the form (4.51), is symmetric positive semi-definite.*

Proof. According to (4.51), the blocks $\Gamma(\mathbf{u}_i)$ with index $1 \leq i \leq p$, such that $\|\mathbf{u}_i\| < \frac{1}{\gamma}$, are given by the identity matrices $I_{n_i} \in \mathbb{R}^{n_i \times n_i}$. Therefore, we focus on the remaining blocks, where $\|\mathbf{u}_i\| \geq \frac{1}{\gamma}$. In this case, we have that, for any $\mathbf{z}_i \in \mathbb{R}^{n_i}$:

$$\langle \mathbf{z}_i, \Gamma(\mathbf{u}_i)\mathbf{z}_i \rangle = \left\langle \mathbf{z}_i, \left(\frac{1}{\|\mathbf{u}_i\|} I_{n_i} - \frac{\mathbf{u}_i \mathbf{u}_i^\top}{\|\mathbf{u}_i\|^3} \right) \mathbf{z}_i \right\rangle = \frac{\|\mathbf{z}_i\|^2}{\|\mathbf{u}_i\|} - \left\langle \mathbf{z}_i, \frac{\mathbf{u}_i \mathbf{u}_i^\top}{\|\mathbf{u}_i\|^3} \mathbf{z}_i \right\rangle = \frac{\|\mathbf{z}_i\|^2}{\|\mathbf{u}_i\|} - \frac{\langle \mathbf{z}_i, \mathbf{u}_i \rangle^2}{\|\mathbf{u}_i\|^3}. \tag{4.52}$$

Using the Cauchy-Schwarz inequality in the second term in (4.52) we get that

$$\langle \mathbf{z}_i, \Gamma(\mathbf{u}_i)\mathbf{z}_i \rangle \quad \geq \frac{\|\mathbf{z}_i\|^2}{\|\mathbf{u}_i\|} - \frac{\|\mathbf{z}_i\|^2 \|\mathbf{u}_i\|^2}{\|\mathbf{u}_i\|^3} = 0.$$

By collecting these results we get that each i-th block $\Gamma(\mathbf{u}_i)$, for $1 \leq i \leq p$, is symmetric positive semi-definite. Therefore, $\Gamma(\mathbf{u})$ is symmetric positive semi-definite because a block-diagonal matrix is positive semi-definite, if and only if, each diagonal block is positive semi-definite. □

By combining the Hessian approximation of the regular part and the generalized second-order information for the $\|\cdot\|_{1,2}$ norm, given by $\Gamma(\mathbf{u})$, the second-order matrix we are going to use is given by

$$H(\mathbf{u}^k) = (B(\mathbf{u}^k) + \sigma \Gamma(\mathbf{u}^k)) \in \mathbb{R}^{m \times m}.$$

Reduced second-order matrix

In Section 4.3.3, we have anticipated that second-order information will be used in the inactive groups to expedite the descent process. By constructing a reduced matrix from this submatrix and multiplying it by the descent directions, the iterations can be accelerated, similar to what occurs in optimization problems with box constraints. Second-order information is useless for the *predicted* active indices $i \in \tilde{\mathcal{A}}^k = \mathcal{A}_0^k \cup \mathcal{A}_\nabla^k \cup \mathcal{J}^k$. For instance, if the i-th group belongs to the active set $\mathcal{A}^k$, it fulfills the necessary optimality condition, implying that its descent direction is the zero vector and eliminating the need for modification. Likewise, if the i-th group is categorized under set $\mathcal{A}_\nabla^k$, it does no satisfy a necessary optimality condition. Therefore, the descent procedure is performed along the steepest descent direction $-\bar{\mathbf{v}}_i^k$. In this case second-order information does not provide curvature information.

Accordingly, let us define the reduced second-order matrix denoted by $H_R(\mathbf{u}^k)$ as follows:

$$[H_R(\mathbf{u}^k)]_{j,l} = \begin{cases} D_{j,l} & \text{if } j \in \tilde{\mathcal{A}}^k \text{ or } l \in \tilde{\mathcal{A}}^k \\ [B(\mathbf{u}^k) + \sigma H(\mathbf{u}^k)]_{j,l} & \text{otherwise.} \end{cases} \tag{4.53}$$

where $D_{j,l}$ denotes a generalization of the Kronecker delta given by:

$$D_{j,l} = \begin{cases} I_{n_j} & \text{if } j = l \\ \mathbf{0}_{n_j,n_l} & \text{if } j \neq l, \end{cases} \tag{4.54}$$

and $\mathbf{0}_{n_j,n_l}$ denotes a matrix with zero entries of size $n_j \times n_l$.

Henceforth, we omit the dependence on the argument in the matrix operators. For simplicity we write H_R^k instead of $H_R(\mathbf{u}^k)$.

Next, we modify the search direction $-\tilde{\mathbf{d}}^k$ and propose a second-order descent direction by integrating the curvature information exclusively in the inactive groups. Accordingly, we solve the following Newton-like system at each k-th iteration:

$$H_R^k \mathbf{w}^k = -\tilde{\mathbf{d}}^k. \tag{4.55}$$

By reordering the elements of $-\tilde{\mathbf{d}}^k$ in such a way that the groups indexed in the inactive set $\tilde{\mathcal{I}}^k$ appear first, the reduced system (4.55) can be rewritten as follows:

$$\begin{pmatrix} \left(B^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]} + \sigma\Gamma^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}\right) & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \begin{pmatrix} \mathbf{w}^k_{\tilde{\mathcal{I}}^k} \\ \mathbf{w}^k_{\tilde{\mathcal{A}}^k} \end{pmatrix} = - \begin{pmatrix} \tilde{\mathbf{d}}^k_{\tilde{\mathcal{I}}^k} \\ \tilde{\mathbf{d}}^k_{\tilde{\mathcal{A}}^k} \end{pmatrix}, \tag{S}$$

where $\tilde{\mathbf{d}}^k_{\tilde{\mathcal{I}}^k}$ and $\tilde{\mathbf{d}}^k_{\tilde{\mathcal{A}}^k}$ correspond to the portions of vector $\tilde{\mathbf{d}}^k$ containing all the groups

indexed in set $\tilde{\mathcal{I}}^k$ and $\tilde{\mathcal{A}}^k$, respectively. The same characterization follows for $\mathbf{w}^k_{\tilde{\mathcal{I}}^k}$ and $\mathbf{w}^k_{\tilde{\mathcal{A}}^k}$.

Likewise, $\left(B^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]} + \sigma\Gamma^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}\right)$ is the square submatrix of H^k with elements associated with the inactive index-set $\tilde{\mathcal{I}}^k$.

Assumption 4.1 (Hessian properties)**.** *Let us assume that, at a local minimizer* $\mathbf{u}^*$*, the reduced Hessian of the regular part* $\nabla^2 f(\mathbf{u}^*)_{[\mathcal{I}^*,\mathcal{I}^*]}$ *is a positive definite matrix. Moreover, we assume that for each iterate* $\mathbf{u}^k$*, the reduced Hessian approximating matrix* H^k_R *satisfies*

$$c\|\mathbf{x}\|^2 \leq \langle \mathbf{x}, H^k_R\mathbf{x}\rangle \leq C\|\mathbf{x}\|^2, \tag{4.56}$$

for all $\mathbf{x} \in \mathbb{R}^m \setminus \{\mathbf{0}\}$*, and some positive constants* c *and* C*, independent of* k*.*

The second order linear system (**S**) is therefore solvable, and we can obtain the descent direction $\mathbf{w}^k$ by solving it. The next theorem shows that this direction is a descent direction for the function ψ.

Theorem 4.4. *Let* $\mathbf{w}^k \neq 0$ *be the solution of system* (**S**)*. Then,* $\mathbf{w}^k$ *satisfies*

$$\psi'(\mathbf{u}^k, \mathbf{w}^k) \leq -c\|\mathbf{w}^k_{\tilde{\mathcal{I}}^k}\|^2 - \sum_{i\in\mathcal{A}^k_\nabla} \|\bar{\mathbf{v}}^k_i\|^2 < 0. \tag{4.57}$$

Therefore, $\mathbf{w}^k$ *is a descent direction.*

Proof. The directional derivative of ψ at $\mathbf{u}^k$ in the direction $\mathbf{w}^k$ is given by

$$\psi'(\mathbf{u}^k, \mathbf{w}^k) = \langle \nabla f(\mathbf{u}^k), \mathbf{w}^k\rangle + \sigma \sum_{i=1}^{p} h'(\mathbf{u}^k_i, \mathbf{w}^k).$$

Taking into account the *predicted* active and inactive index-sets $\tilde{\mathcal{A}}^k$ and $\tilde{\mathcal{I}}^k$, recalling that $\tilde{\mathcal{A}}^k = \mathcal{A}^k_0 \cup \mathcal{A}^k_\nabla \cup \mathcal{J}^k$, and considering whether $\mathbf{u}^k_i$ is zero or nonzero, from the directional derivative of the Euclidean norm given in (4.35), we get:

$$\begin{aligned}\psi'(\mathbf{u}^k, \mathbf{w}^k) \;&= \sum_{i\in\tilde{\mathcal{I}}^k} \left\langle \nabla_i f(\mathbf{u}^k) + \sigma\frac{\mathbf{u}^k_i}{\|\mathbf{u}^k_i\|}, \mathbf{w}^k_i\right\rangle + \sum_{i\in\mathcal{A}^k_0\cup\mathcal{A}^k_\nabla} \langle \nabla_i f(\mathbf{u}^k), \mathbf{w}^k_i\rangle + \sigma\|\mathbf{w}^k_i\| \\ &\quad + \sum_{i\in\mathcal{J}^k} \left\langle \nabla_i f(\mathbf{u}^k) + \sigma\frac{\mathbf{u}^k_i}{\|\mathbf{u}^k_i\|}, \mathbf{w}^k_i\right\rangle.\end{aligned} \tag{4.58}$$

We analyze each sum on the right-hand side of (4.58) separately.

From system (**S**) and (4.44) we know that

$$\left(B^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]} + \sigma\Gamma^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}\right)\mathbf{w}^k_{\tilde{\mathcal{I}}^k} = -\,\bar{\mathbf{v}}^k_{\tilde{\mathcal{I}}^k}.$$

Hence, from the definition of $-\bar{\mathbf{v}}^k$ in (4.38) we get that the sum over the *predicted* index-set $\tilde{\mathcal{I}}^k$ in (4.58) is given by:

$$\sum_{i\in\tilde{\mathcal{I}}^k}\left\langle \nabla_i f(\mathbf{u}^k)+\sigma\frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|},\mathbf{w}_i^k\right\rangle=\sum_{i\in\tilde{\mathcal{I}}^k}\langle\bar{\mathbf{v}}_i^k,\mathbf{w}_i^k\rangle \;=\langle\bar{\mathbf{v}}_{\tilde{\mathcal{I}}^k}^k,\mathbf{w}_{\tilde{\mathcal{I}}^k}^k\rangle$$
$$=-\left\langle\mathbf{w}_{\tilde{\mathcal{I}}^k}^k,\left(B_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}^k+\sigma\Gamma_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}^k\right)\mathbf{w}_{\tilde{\mathcal{I}}^k}^k\right\rangle.$$

Additionally, from Assumption 4.1 and the symmetric positive definiteness of the sub-matrix $\Gamma_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}^k$ (see Proposition 4.2) we have that

$$\begin{aligned}\sum_{i\in\tilde{\mathcal{I}}^k}\left\langle \nabla_i f(\mathbf{u}^k)+\sigma\frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|},\mathbf{w}_i^k\right\rangle &=-\left\langle\mathbf{w}_{\tilde{\mathcal{I}}^k}^k,\left(B_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}^k+\sigma\Gamma_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}^k\right)\mathbf{w}_{\tilde{\mathcal{I}}^k}^k\right\rangle\\ &=-\left\langle\mathbf{w}_{\tilde{\mathcal{I}}^k}^k,B_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}^k\mathbf{w}_{\tilde{\mathcal{I}}^k}^k\right\rangle-\sigma\left\langle\mathbf{w}_{\tilde{\mathcal{I}}^k}^k,\Gamma_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}^k\mathbf{w}_{\tilde{\mathcal{I}}^k}^k\right\rangle\\ &<-\tilde{c}\|\mathbf{w}_{\tilde{\mathcal{I}}^k}^k\|^2.\end{aligned}\tag{4.59}$$

For the second sum in (4.58), thanks to (4.44) and the structure of the reduced matrix (4.53), the components of the steepest descent direction remain unchanged, i.e., $-\tilde{\mathbf{d}}_i^k=-\bar{\mathbf{v}}_i^k$, for all $i\in\mathcal{A}_0^k\cup\mathcal{A}_\nabla^k$. Moreover, from system (S), we deduce that $\mathbf{w}_{\mathcal{A}_0^k\cup\mathcal{A}_\nabla^k}^k=-\tilde{\mathbf{d}}_{\mathcal{A}_0^k\cup\mathcal{A}_\nabla^k}^k=-\bar{\mathbf{v}}_{\mathcal{A}_0^k\cup\mathcal{A}_\nabla^k}^k$. From this result and since $\bar{\mathbf{v}}_i^k=0$ for all $i\in\mathcal{A}_0^k$, the second sum in (4.58) becomes:

$$\begin{aligned}\sum_{i\in\mathcal{A}_0^k\cup\mathcal{A}_\nabla^k}\langle\nabla_i f(\mathbf{u}^k),\mathbf{w}_i^k\rangle+\sigma\|\mathbf{w}_i^k\| &=\sum_{i\in\mathcal{A}_0^k\cup\mathcal{A}_\nabla^k}\langle\nabla_i f(\mathbf{u}^k),-\bar{\mathbf{v}}_i^k\rangle+\sigma\|\bar{\mathbf{v}}_i^k\|\\ &=\sum_{i\in\mathcal{A}_\nabla^k}\langle\nabla_i f(\mathbf{u}^k),-\bar{\mathbf{v}}_i^k\rangle+\sigma\|\bar{\mathbf{v}}_i^k\|\\ &=\sum_{i\in\mathcal{A}_\nabla^k}\left\langle\nabla_i f(\mathbf{u}^k),-\nabla_i f(\mathbf{u}^k)+\sigma\frac{\nabla_i f(\mathbf{u}^k)}{\|\nabla_i f(\mathbf{u}^k)\|}\right\rangle\\ &+\sigma\left\|-\nabla_i f(\mathbf{u}^k)+\sigma\tfrac{\nabla_i f(\mathbf{u}^k)}{\|\nabla_i f(\mathbf{u}^k)\|}\right\|.\end{aligned}$$

Proceeding as in the proof of Proposition 4.1, see equations (4.49) and (4.50), we obtain that

$$\sum_{i\in\mathcal{A}_0^k\cup\mathcal{A}_\nabla^k}\langle\nabla_i f(\mathbf{u}^k),\mathbf{w}_i^k\rangle+\sigma\|\mathbf{w}_i^k\| \;=-\sum_{i\in\mathcal{A}_\nabla^k}\|\bar{\mathbf{v}}_i^k\|^2.\tag{4.60}$$

Analogously, from system (S) and (4.44), we obtain that $\mathbf{w}_{\mathcal{J}^k}^k=-\tilde{\mathbf{d}}_{\mathcal{J}^k}^k=-\mathbf{u}_{\mathcal{J}^k}^k$. Together with (4.38), it follows that the third term in (4.58) can be expressed as:

$$\sum_{i\in\mathcal{J}^k}\left\langle\nabla_i f(\mathbf{u}^k)+\sigma\frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|},\mathbf{w}_i^k\right\rangle=\sum_{i\in\mathcal{J}^k}\langle\bar{\mathbf{v}}_i^k,-\mathbf{u}_i^k\rangle<0.\tag{4.61}$$

Finally, by replacing (4.59), (4.60) and (4.61) into (4.58), we obtain the result. □

The following assumption allows us to improve the last result and is required for the global convergence properties of the algorithm. Also, from a practical point of view, it ensures a more robust projection step.

Assumption 4.2. *There exists $\epsilon > 0$ such that the index-set $\mathcal{J}^k$ used in the active-set prediction is given by*

$$\mathcal{J}^k = \{i : \langle \mathbf{u}_i^k, -\mathbf{v}_i^k \rangle \leq -\epsilon \|\mathbf{u}_i^k\|^2\}$$

Corollary 4.1. *Under Assumptions 4.1 and 4.2 the descent direction $\mathbf{w}^k$ in Theorem 4.4 satisfies*

$$\psi^{'}(\mathbf{u}^k, \mathbf{w}^k) \leq -c\|\mathbf{w}^k\|_{2,1}^2. \tag{4.62}$$

for some positive constant c.

Proof. Thanks to our assumptions, the relation (4.61) in the proof of Theorem 4.4 becomes

$$\sum_{i \in \mathcal{J}^k} \left\langle \nabla_i f(\mathbf{u}^k) + \sigma \frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}, \mathbf{w}_i^k \right\rangle = \sum_{i \in \mathcal{J}^k} \langle \bar{\mathbf{v}}_i^k, -\mathbf{u}_i^k \rangle < -\epsilon \sum_{i \in \mathcal{J}^k} \|\mathbf{u}_i^k\|^2 = -\epsilon \sum_{i \in \mathcal{J}^k} \|\mathbf{w}_i^k\|^2. \tag{4.63}$$

Then, (4.59), (4.60), and (4.63) imply the result. □

4.3.4 GSDM Algorithm

The Group-Sparse Active-Set Descent Method (GSDM) is built upon the descent direction $\mathbf{w}^k$, solution of (**S**), which needs to be complemented with a line-search strategy for the computation of the descent step. The line-search shall provide conditions that guarantee descent of the cost function.

We utilize a generalized Armijo condition tailored for locally Lipschitz continuous functions that are not necessarily convex [121]. For details on the sufficient decrease condition we refer the reader to Section 4.6. Based on the analysis of the previous section we consolidate the described method in the following algorithm.

Algorithm 5: GSDM

Initialize $\mathbf{u}^0$, set $k = 0$ and tol > 0 ;
while *stopping criterion is not satisfied* **do**
  for $i \leftarrow 1$ **to** p **do**
    Compute the following index-sets;
    $\mathcal{A}_0^k := \{i : \|\mathbf{u}_i^k\| = 0 \text{ and } \|\nabla_i f(\mathbf{u}^k)\| \leq \sigma\}$;
    $\mathcal{A}_\nabla^k := \{i : \|\mathbf{u}_i^k\| = 0 \text{ and } \|\nabla_i f(\mathbf{u}^k)\| > \sigma\}$;
    $\mathcal{I}^k := \{i : \|\mathbf{u}_i^k\| \neq 0\}$;
    compute vector $-\bar{\mathbf{v}}_i^k$ given by;

$$\bar{\mathbf{v}}_i^k = \begin{cases} \nabla_i f(\mathbf{u}^k) + \sigma \frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}, & \text{if } i \in \mathcal{I}^k, \\ \mathbf{0}, & \text{if } i \in \mathcal{A}_0^k, \\ \nabla_i f(\mathbf{u}^k) - \sigma \frac{\nabla_i f(\mathbf{u}^k)}{\|\nabla_i f(\mathbf{u}^k)\|}, & \text{if } i \in \mathcal{A}_\nabla^k. \end{cases}$$

  end
  if $k > 1$ **then**
    for $i \in \mathcal{I}^k$ **do**
      if $\langle \mathbf{u}_i^k, -\bar{\mathbf{v}}_i^k \rangle < tol$ **then**
        store i in $\mathcal{J}^k$;
        set $-\bar{\mathbf{v}}_i^k \leftarrow -\mathbf{u}_i^k$;
                                                {Active-set phase}
      end
    end
  end
  set $\mathbf{w}^k \leftarrow -\bar{\mathbf{v}}^k$;
  compute $\tilde{\mathcal{I}}^k = \mathcal{I}^k \setminus \mathcal{J}^k$;
  compute $\left(B^k_{[\tilde{\mathcal{I}}^k, \tilde{\mathcal{I}}^k]} + \sigma \Gamma^k_{[\tilde{\mathcal{I}}^k, \tilde{\mathcal{I}}^k]}\right)$, with Γ^k given in (4.51);
  solve $\left(B^k_{[\tilde{\mathcal{I}}^k, \tilde{\mathcal{I}}^k]} + \sigma \Gamma^k_{[\tilde{\mathcal{I}}^k, \tilde{\mathcal{I}}^k]}\right) \mathbf{x}^k = -\bar{\mathbf{v}}_{\tilde{\mathcal{I}}^k}$;
  set $\mathbf{w}^k_{\tilde{\mathcal{I}}^k} \leftarrow \mathbf{x}^k$;
  execute a line-search routine to determine the step size s^k;
  $\mathbf{u}^k \leftarrow \mathbf{u}^k + s^k \mathbf{w}^k$;
  $k \leftarrow k + 1$;
end

In the next theorem we provide a global convergence result for our algorithm based on the theory developed in [94], which in particular requires the function f to satisfy the Kurdyka-Łojasiewicz property (KL).

Theorem 4.5. *Suppose that Assumptions 4.1 and 4.2 hold and that ψ satisfies the following Kurdyka-Łojasiewicz property: there exist positive constants κ, ζ and $\theta \in [0, 1[$*

such that for all $\boldsymbol{\xi} \in \sigma\partial(\|\cdot\|_{2,1})(\mathbf{u})$ *and every* $\mathbf{u}$ *in the level set* $\{\mathbf{u} : \psi(\mathbf{u}) \leq \psi(\mathbf{u}_0)\}$*, it holds that*

$$\kappa|\psi(\mathbf{u}) - \psi^*|^\theta \leq \|\nabla f(\mathbf{u}) + \boldsymbol{\xi}\|_{1,2}, \tag{4.64}$$

for $\psi^* \in \mathbb{R}$ *such that* $|\psi(\mathbf{u}) - \psi^*| \leq \zeta$*. Then, the sequence* $\{\mathbf{u}^k\}$ *generated by GSDM (with an Armijo line-search) converges to a point* $\mathbf{u}^*$ *such that* $0 \in \nabla f(\mathbf{u}^*) + \sigma\partial(\|\mathbf{u}^*\|_{1,2})$*.*

Proof. We provide a sketch of the proof since the result is rather standard. The strict descent obtained in Corollary 4.1 and boundedness from below of ψ imply that an Armijo line-search yields a monotone decrease of the sequence $\{\psi(\mathbf{u}^k)\}$. Thus, $\{\psi(\mathbf{u}^k)\}$ converges to some limit ψ^*.

Using c as a generic constant, we deduce from Corollary 4.1 and the KŁ property that

$$\begin{aligned}
\|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{1,2} =& s^k \|\mathbf{w}^k\|_{1,2} \\
\leq& -\frac{s^k}{c}\psi'(\mathbf{u}^k, \mathbf{w}^k) \\
\leq& \frac{t}{c}(\psi(\mathbf{u}^k) - \psi(\mathbf{u}^{k+1})) \\
=& \frac{s^k}{c}(\psi(\mathbf{u}^k) - \psi^* - (\psi(\mathbf{u}^{k+1}) - \psi^*)) \\
\leq& \frac{s^k}{c}\frac{1}{1-\theta}(\psi(\mathbf{u}^k) - \psi^*)^\theta((\psi(\mathbf{u}^k) - \psi^*)^{1-\theta} - (\psi(\mathbf{u}^{k+1}) - \psi^*)^{1-\theta}) \\
\leq& s^k \|\nabla f(\mathbf{u}^k) + \bar{\boldsymbol{\xi}}^k\|_{1,2}\Delta^k,
\end{aligned}$$

where $\bar{\boldsymbol{\xi}}^k \in \sigma\partial\|\cdot\|_{1,2}(\mathbf{u}^k)$ is taken such that $\bar{\boldsymbol{\xi}}_i^k = -\nabla_i f(\mathbf{u}^k)$ for $i \in \mathcal{A}_0^k$. $\Delta^k := \frac{c}{1-\theta}((\psi(\mathbf{u}^k) - \psi^*)^{1-\theta} - (\psi(\mathbf{u}^{k+1}) - \psi^*)^{1-\theta}$. From the definition of $\tilde{\mathbf{d}}^k$ and the system (S) we have

$$\|\nabla f(\mathbf{u}^k) + \bar{\boldsymbol{\xi}}^k\|_{1,2} = \sum_{i\in\tilde{\mathcal{I}}} \|\tilde{\mathbf{d}}_i^k\| + \sum_{i\in\mathcal{A}_\nabla^k} \|\tilde{\mathbf{d}}_i^k\| + \underbrace{\sum_{i\in\mathcal{A}_0^k} \|\tilde{\mathbf{d}}_i^k\|}_{=0} + \sum_{i\in\mathcal{J}^k} \|\bar{\mathbf{v}}_i^k\|$$

Taking into account Assumption 4.1 in the first sum, and Assumption (4.2) in the last sum we arrive to

$$\|\nabla f(\mathbf{u}^k) + \bar{\boldsymbol{\xi}}^k\|_{1,2} \leq c\sum_{i=1}^{p} \|\mathbf{w}_i^k\|,$$

for some constant $c > 0$. Thus, taking $s^k \leq \bar{s}$, we obtain the estimate

$$\|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{1,2} \leq \bar{s}c\|\mathbf{w}^k\|_{1,2}\Delta^k.$$

It remains to verify the summability of $\{\|\mathbf{w}^k\|_{1,2}\}$. This is obtained by [29, Lemma 4.3] with Δ^k (summable by [29, Theorem 4.1]), $u^k = \|\mathbf{w}^k\|_{1,2}$, $g^k = \psi(\mathbf{u}^k) - \psi^*$ and $\hat{g}^k = \psi(\mathbf{u}^{k+1}) - \psi^*$. The summability of $\{\|\mathbf{w}^k\|\}$ implies that $\{\|\mathbf{u}^{k+1}\|_{1,2}\}$ is a Cauchy sequence, hence its convergence to a limit $\mathbf{u}^*$. By the properties of the graph of $\nabla f + \sigma\partial\|\cdot\|_{1,2}$ we conclude that $\mathbf{u}^*$ is a critical point. □

4.4 Identification of the Active and Inactive Sets

In this section, we investigate how the different active and inactive sets of our algorithm evolve along the iterations. In particular, we show that the active and inactive sets of a local minimizer $\mathbf{u}^*$ are identified once the iterations reach a certain neighborhood of $\mathbf{u}^*$, i.e.,

$$\tilde{\mathcal{I}}^k = \mathcal{I}^k = \mathcal{I}^*, \quad \text{and} \quad \tilde{\mathcal{A}}^k = \mathcal{A}_0^k = \mathcal{A}^*, \quad \text{for all } k > k_0,$$

for some $k_0 \in \mathbb{N}$. This important result enables to obtain a superlinear convergence rate of the algorithm iterates to the solution $\mathbf{u}^*$, as will be shown in Section 4.5.

Let us denote the active and inactive index sets at the solution $\mathbf{u}^*$, denoted by $\mathcal{A}^*$ and $\mathcal{I}^*$, respectively, are defined as

$$\mathcal{A}^* := \{i : \|\mathbf{u}_i^*\| = 0\}, \quad \text{and} \quad \mathcal{I}^* := \{i : \|\mathbf{u}_i^*\| > 0\}.$$

To support this analysis, we introduce the following strict complementarity assumption related to the necessary optimality condition (4.32).

Assumption 4.3 (Strict complementarity)**.** *Let $\mathbf{u}^*$ be a local minimizer such that it satisfies*

$$\begin{aligned} \|\nabla f(\mathbf{u}^*)\| &< \sigma, && \text{if } \|\mathbf{u}_i^*\| = 0, \\ -\frac{1}{\sigma}\nabla f(\mathbf{u}^*) &= \frac{\mathbf{u}_i^*}{\|\mathbf{u}_i^*\|}, && \text{if } \|\mathbf{u}_i^*\| > 0. \end{aligned}$$

The following result provides a useful local characterization of the active and inactive index sets in the vicinity of a local minimizer $\mathbf{u}^*$.

Theorem 4.6. *Let ∇f be Lipschitz continuous with constant L. Suppose $\mathbf{u}^*$ is a local minimizer of problem (GS), satisfying Assumption 4.3. Let $\mathbf{u}^k$ be the k-th iteration generated by GSDM, with $\mathbf{u}^k \in B(\mathbf{u}^*, r)$. If r is chosen sufficiently small, such that $0 < r < \min\left\{\min_{i\in\mathcal{I}^*}\|\mathbf{u}_i^*\|, \frac{\sigma - \max_{i\in\mathcal{A}^*}\|\nabla_i f(\mathbf{u}^*)\|}{L}\right\}$, then the following results hold:*

i) $\tilde{\mathcal{I}}^k \subset \mathcal{I}^*$,

ii) $\mathcal{A}_0^k \subset \mathcal{A}^*$,

iii) $\mathcal{A}_{\nabla}^{k} = \emptyset$,

iv) $\mathcal{I}^{*} \subset \mathcal{I}^{k}$.

Proof. i) We proceed by contradiction. Let us suppose there exists at least one index $i \in \tilde{\mathcal{I}}^{k}$ such that $i \notin \mathcal{I}^{*}$, meaning $i \in \mathcal{A}^{*}$. From $i \in \tilde{\mathcal{I}}^{k} = \mathcal{I}^{k} \setminus \mathcal{J}^{k}$ we get that $0 \leq \langle \mathbf{u}_i^k, -\bar{\mathbf{v}}_i^k \rangle = \langle \mathbf{u}_i^k, -\nabla_i f(\mathbf{u}^k) \rangle - \langle \mathbf{u}_i^k, \sigma \frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|} \rangle$, which leads to

$$\sigma \leq \frac{1}{\|\mathbf{u}_i^k\|} \langle \mathbf{u}_i^k, -\nabla_i f(\mathbf{u}^k) \rangle \leq \|\nabla_i f(\mathbf{u}^k)\|. \tag{4.65}$$

Now, from (4.65) and the Lipschitz continuity of the gradient ∇f at $\mathbf{u}^k \in B(\mathbf{u}^*, r)$, we have

$$\begin{aligned} \sigma \leq \|\nabla_i f(\mathbf{u}^k) - \nabla_i f(\mathbf{u}^*) + \nabla_i f(\mathbf{u}^*)\| &\leq L\|\mathbf{u}^k - \mathbf{u}^*\| + \|\nabla_i f(\mathbf{u}^*)\| \\ &\leq Lr + \|\nabla_i f(\mathbf{u}^*)\|. \end{aligned} \tag{4.66}$$

According to the strict complementary Assumption 4.3, we have $\|\nabla_i f(\mathbf{u}^*)\| < \sigma$ for all $i \in \mathcal{A}^*$. In particular, $\max_{i \in \mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\| < \sigma$. Then, by taking $r < \frac{\sigma - \max_{i \in \mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L}$ in (4.66) we get the following contradiction:

$$\sigma < \sigma - \max_{i \in \mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\| + \|\nabla_i f(\mathbf{u}^*)\| \leq \sigma.$$

Therefore, we conclude that $\tilde{\mathcal{I}}^{k} \subset \mathcal{I}^{*}$.

ii) We proceed by contradiction. Let us suppose there exists at least one index $i \in \mathcal{A}_0^k$ such that $i \notin \mathcal{A}^*$, meaning $0 < \|\mathbf{u}_i^*\|$. On the other hand, since $\|\mathbf{u}_i^k\| = 0$, we have that

$$\|\mathbf{u}_i^*\| = \|\mathbf{u}_i^* - \mathbf{u}_i^k\| \leq \|\mathbf{u}^* - \mathbf{u}^k\| < r.$$

Now, given that $r < \min\left\{\min_{i \in \mathcal{I}^*} \|\mathbf{u}_i^*\|, \frac{\sigma - \max_{i \in \mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L}\right\}$, it follows that

$$0 < \|\mathbf{u}_i^*\| < r < \min_{i \in \mathcal{I}^*} \|\mathbf{u}_i^*\|.$$

Which is a contradiction. Then $i \in \mathcal{A}^*$.

iii) Suppose there exists at least one index $i \in \mathcal{A}_{\nabla}^{k}$. We will proof that both sets $i \in \mathcal{A}_{\nabla}^{k} \cap \mathcal{A}^*$ and $i \in \mathcal{A}_{\nabla}^{k} \cap \mathcal{I}^*$ are empty.

First, consider the case $i \in \mathcal{A}_{\nabla}^{k} \cap \mathcal{A}^*$. We have $\sigma < \|\nabla_i f(\mathbf{u}^k)\|$ and, according to Assumption 4.3, $\|\nabla_i f(\mathbf{u}^*)\| < \sigma$. From the Lipschitz continuity of the gradient ∇f at $\mathbf{u}^k \in B(\mathbf{u}^*, r)$, as in (4.66), we have:

$$\sigma \leq \|\nabla_i f(\mathbf{u}^k) - \nabla_i f(\mathbf{u}^*) + \nabla_i f(\mathbf{u}^*)\| \leq Lr + \|\nabla_i f(\mathbf{u}^*)\|. \tag{4.67}$$

Similarly as in case i), by taking $r < \frac{\sigma - \max_{i\in\mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L}$ in (4.67), we get a contradiction. Therefore, $\mathcal{A}_\nabla^k \cap \mathcal{A}^* = \emptyset$.

We consider the remaining case, $i \in \mathcal{A}_\nabla^k \cap \mathcal{I}^*$, meaning that $\|\mathbf{u}_i^*\| > 0$. By the reverse triangle inequality, it follows that $\|\mathbf{u}_i^*\| - \|\mathbf{u}_i^k\| \leq \|\mathbf{u}_i^* - \mathbf{u}_i^k\| \leq \|\mathbf{u}^* - \mathbf{u}^k\| < r$. Given that $\|\mathbf{u}_i^k\| = 0$, we get $\|\mathbf{u}_i^*\| < r$. Taking $r < \min\{\min_{i\in\mathcal{I}^*} \|\mathbf{u}_i^*\|, \frac{\sigma - \max_{i\in\mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L}\}$ in the last relation, we obtain that $0 < \|\mathbf{u}_i^*\| < \min_{i\in\mathcal{I}^*} \|\mathbf{u}_i^*\|$, which is a contradiction. Thus, $\mathcal{A}_\nabla^k \cap \mathcal{I}^* = \emptyset$.

iv) Let $i \in \mathcal{I}^*$. By the reverse triangle inequality, we have

$$\|\mathbf{u}_i^*\| - \|\mathbf{u}_i^k\| \leq \|\mathbf{u}_i^* - \mathbf{u}_i^k\| \leq \|\mathbf{u}^* - \mathbf{u}^k\| < r.$$

Rearranging the first inequality, we obtain $\|\mathbf{u}_i^*\| - r < \|\mathbf{u}_i^k\|$. Now, given that $r < \min\left\{\min_{i\in\mathcal{I}^*} \|\mathbf{u}_i^*\|, \frac{\sigma - \max_{i\in\mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L}\right\}$, it follows that

$$0 \leq \|\mathbf{u}_i^*\| - \min_{i\in\mathcal{I}^*} \|\mathbf{u}_i^*\| < \|\mathbf{u}_i^k\|.$$

Thus, we conclude that $i \in \mathcal{I}^k$.

□

In the following result, we demonstrate that, if three consecutive iterations remain within a ball of sufficiently small radius, the index set $\mathcal{J}^k$ becomes empty and, consequently, the index-sets $\tilde{\mathcal{I}}^k$ and $\tilde{\mathcal{A}}^k$ coincide with the active and inactive index sets at $\mathbf{u}^*$, respectively.

Theorem 4.7. *Let ∇f be Lipschitz continuous with constant L. Suppose $\mathbf{u}^*$ is a local minimizer satisfying Assumption 4.3. Let $\mathbf{u}^{k-1}$, $\mathbf{u}^k$ and $\mathbf{u}^{k+1}$ be three consecutive full-step iterations generated by the GSDM Algorithm with full steps satisfying $\mathbf{u}^{k-1}, \mathbf{u}^k, \mathbf{u}^{k+1} \in B(\mathbf{u}^*, r)$, where $0 < r < \min\left\{\min_{i\in\mathcal{I}^*} \|\mathbf{u}_i^*\|, \frac{\sigma - \max_{i\in\mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L}\right\}$. Then, the index set $\mathcal{J}^k$ vanishes, and the following equalities hold:*

$$\tilde{\mathcal{I}}^k = \mathcal{I}^k = \mathcal{I}^* \quad \text{and} \quad \tilde{\mathcal{A}}^k = \mathcal{A}_0^k = \mathcal{A}^*.$$

Proof. We begin by proving that $\mathcal{J}^k$ vanishes. We proceed by contradiction and assume that $\mathcal{J}^k \neq \emptyset$. Since $\mathbf{u}^{k-1} \in B(\mathbf{u}^*, r)$, it follows from Theorem 4.6 - iii) that $\mathcal{A}_\nabla^{k-1} = \emptyset$. Consequently, the indices in $\mathcal{J}^k$ must originate from the index set $\tilde{\mathcal{I}}^{k-1}$, meaning that

$$\mathcal{J}^k \subset \tilde{\mathcal{I}}^{k-1} \subset \mathcal{I}^*.$$

Similarly, since $\mathbf{u}^{k+1} \in B(\mathbf{u}^*, r)$, we also have $\mathcal{A}_\nabla^{k+1} = \emptyset$. Thus, during the active set prediction phase the groups indexed in $\mathcal{J}^k$ become sparse at iterate $k+1$. Therefore,

the set $\mathcal{J}^k$ must satisfy

$$\mathcal{J}^k \subset \mathcal{A}_0^{k+1} \subset \mathcal{A}^*.$$

Since $\mathcal{I}^*$ and $\mathcal{A}^*$ are disjoint index sets, we obtain that $\mathcal{J}^k$ must be empty.

Next, given that $\mathcal{J}^k$ is empty, from the definition of the index set $\tilde{\mathcal{A}}^k$ we have $\tilde{\mathcal{A}}^k = \mathcal{A}_\nabla^k \cup \mathcal{A}_0^k$. However, since $\mathbf{u}^k \in B(\mathbf{u}^*, r)$, it follows from Theorem 4.6 that $\mathcal{A}_\nabla^k$ vanishes. Thus, we obtain $\tilde{\mathcal{A}}^k = \mathcal{A}_0^k$. On the other hand, by definition, we have $\tilde{\mathcal{I}}^k = \mathcal{I}^k \setminus \mathcal{J}^k$. Since $\mathcal{J}^k$ vanishes, it follows that $\tilde{\mathcal{I}}^k = \mathcal{I}^k$. Moreover, applying Theorem 4.6- i) and iv), we obtain $\tilde{\mathcal{I}}^k = \mathcal{I}^k = \mathcal{I}^*$. Finally, since $\mathcal{A}^*$ and $\mathcal{I}^*$ are disjoint index-sets, Theorem 4.6 - ii) implies that $\tilde{\mathcal{A}}^k = \mathcal{A}_0^k = \mathcal{A}^*$. □

From Theorems 4.6 and 4.7, we can directly derive the following corollary.

Corollary 4.2. *Let ∇f be Lipschitz continuous. If for some $k_0 \in \mathbb{N}$ the sequence $\{\mathbf{u}^k\}_{k \geq k_0}$ remains within the ball $B(\mathbf{u}^*, r)$, where r is defined as in Theorem 4.7 and $\mathbf{u}^*$ satisfies Assumption 4.3, then the active and inactive index-sets are identified as $\tilde{\mathcal{I}}^k = \mathcal{I}^k = \mathcal{I}^*$ and $\tilde{\mathcal{A}}^k = \mathcal{A}_0^k = \mathcal{A}^*$ for all $k > k_0$.*

Proof. The result follows from Theorems 4.6 and 4.7. □

We can obtain a more precise characterization of the index-set $\tilde{\mathcal{I}}^k$ within the ball $B(\mathbf{u}^*, r)$ by refining the radius r, given in Theorem 4.7, as follows.

Proposition 4.3. *Let $\mathbf{u}^k$ be the k-th approximate solution generated by the GSDM Algorithm, satisfying $\mathbf{u}^k \in B(\mathbf{u}^*, r)$, with*

$$r \leq \frac{1}{\gamma} < \min\left\{ \min_{i \in \mathcal{I}^*} \|\mathbf{u}_i^*\|, \frac{\sigma - \max_{i \in \mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L} \right\}.$$

Then the predicted inactive index-set is characterized as

$$\tilde{\mathcal{I}}^k = \{i : \|\mathbf{u}_i^k\| > 1/\gamma\}.$$

Proof. Consider the indices $i \in \tilde{\mathcal{I}}^k$ such that $0 < \|\mathbf{u}_i^k\| < \frac{1}{\gamma}$. We aim to show that this leads to a contradiction. Since $\mathbf{u}^k \in B(\mathbf{u}^*, r)$, applying the reverse triangle inequality gives

$$\|\mathbf{u}_i^*\| - \|\mathbf{u}_i^k\| \leq \|\mathbf{u}_i^* - \mathbf{u}_i^k\| \leq \|\mathbf{u}^* - \mathbf{u}^k\| < r.$$

Since $r \leq \frac{1}{\gamma}$, we obtain

$$\|\mathbf{u}_i^*\| < \frac{1}{\gamma} + \|\mathbf{u}_i^k\| < \frac{2}{\gamma}.$$

Now, choosing $\gamma > \frac{2}{\min_{i \in \mathcal{I}^*} \|\mathbf{u}_i^*\|}$ ensures that $\frac{2}{\gamma} < \min_{i \in \mathcal{I}^*} \|\mathbf{u}_i^*\|$. Since $i \in \tilde{\mathcal{I}}^*$ (by Theorem 4.6), we know that $\|\mathbf{u}_i^*\| > 0$. Combining this with the previous inequality,

we arrive at $0 < \|\mathbf{u}_i^*\| < \min_{i\in\mathcal{I}^*} \|\mathbf{u}_i^*\|$, which is a contradiction. Thus, no such index i can exist, completing the proof. □

With these results we can establish a local convergence analysis in the next section.

4.5 Local Superlinear Convergence

Building on the results established in Section 4.4, we observe that if the hypotheses of Corollary 4.2 hold, the index sets are identified, i.e.,

$$\tilde{\mathcal{I}}^k = \mathcal{I}^k = \mathcal{I}^* \quad \text{and} \quad \tilde{\mathcal{A}}^k = \mathcal{A}_0^k = \mathcal{A}^*, \quad \text{for all } k > k_0, \tag{4.68}$$

for some $k_0 \in \mathbb{N}$. Building on this identification, we now analyze the second-order system (**S**) at the k-th iteration, for $k > k_0$, restricting our focus to the index sets $\mathcal{I}^*$ and $\mathcal{A}^*$. Consequently, system (**S**) can be written as follows:

$$\begin{pmatrix} B^k_{[\mathcal{I}^*,\mathcal{I}^*]} + \sigma\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]} & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \begin{pmatrix} \mathbf{w}^k_{\mathcal{I}^*} \\ \mathbf{w}^k_{\mathcal{A}_0^*} \end{pmatrix} = - \begin{pmatrix} \bar{\mathbf{v}}^k_{\mathcal{I}^*} \\ \mathbf{0}_{\mathcal{A}_0^*} \end{pmatrix}, \tag{4.69}$$

where, thanks to the characterization of $\tilde{\mathcal{I}}^k$ given in Lemma 4.3, the submatrix $\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}$ is block-diagonal with each diagonal block given by:

$$\Gamma_i^k = \frac{1}{\|\mathbf{u}_i^k\|} I_{n_i} - \frac{\mathbf{u}_i^k \mathbf{u}_i^{k^\top}}{\|\mathbf{u}_i^k\|^3}, \quad \text{for all } i \in \mathcal{I}^*. \tag{4.70}$$

Moreover, for the right-hand side, $-\bar{\mathbf{v}}_i^k = -\nabla_i f(\mathbf{u}^k) - \sigma \frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|}$, for all $i \in \mathcal{I}^*$.

In order to prove convergence, we will consider $B(\mathbf{u})_{[\mathcal{I}^*,\mathcal{I}^*]} = \nabla^2 f(\mathbf{u})_{[\mathcal{I}^*,\mathcal{I}^*]}$ in a neighborhood of a local solution $\mathbf{u}^*$. Further, by Assumption 4.1 we have that the submatrix

$$H^*_{R_{[\tilde{\mathcal{I}}^*,\tilde{\mathcal{I}}^*]}} = \nabla^2 f(\mathbf{u}^*)_{[\mathcal{I}^*,\mathcal{I}^*]} + \sigma\Gamma^*_{[\mathcal{I}^*,\mathcal{I}^*]}$$

is invertible. Morevover, since $\nabla^2 f(\mathbf{u})_{[\tilde{\mathcal{I}}^*,\tilde{\mathcal{I}}^*]}$ is locally Lipschitz continuous and the function $\Gamma(\mathbf{u})_{[\tilde{\mathcal{I}}^*,\tilde{\mathcal{I}}^*]}$ is locally Lipschitz continuous as well (see [66]), the matrix function $H_R(\mathbf{u})_{[\tilde{\mathcal{I}}^*,\tilde{\mathcal{I}}^*]}$ is Lipschitz continuous in a neighbourhood of $\mathbf{u}^*$.

As a consequence of the latter, we obtain the following Lemma, whose proof follows by standard arguments [53, Lemma 7.3].

Lemma 4.1. *There exists $\delta > 0$ such that for all $\mathbf{u}^k \in B(\mathbf{u}^*, \delta)$*

$$\|(H^k_{R_{[\mathcal{I}^*,\mathcal{I}^*]}})^{-1}\| = \left\| \left(\nabla^2 f(\mathbf{u}^k)_{[\mathcal{I}^*,\mathcal{I}^*]} + \sigma\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}\right)^{-1} \right\| \leq 2\|(H^*_{R_{[\mathcal{I}^*,\mathcal{I}^*]}})^{-1}\| =: C_H.$$

To achieve local convergence we summaryze the hypothesis of section 4.4 and state the following assumptions:

Assumption 4.4. *Let us suppose that the sequence generated by the GSDM method, $\{\mathbf{u}^k\}_{k\geq k_0}$ for some $k_0 \in \mathbb{N}$, remains within the ball $B(\mathbf{u}^*, r)$, where r satisfies*

$$0 < r < \min\left\{\min_{i\in\mathcal{I}^*} \|\mathbf{u}_i^*\|, \frac{\sigma - \max_{i\in\mathcal{A}^*} \|\nabla_i f(\mathbf{u}^*)\|}{L}, \frac{1}{\gamma}\right\}.$$

We now establish a local convergence result, beginning with the following convergence rate estimate.

Theorem 4.8. *Let $\mathbf{u}^*$ is a local minimizer of problem (GS) such that Assumption 4.4 holds. Then there exist $\hat{L} > 0$ and $\epsilon > 0$ such that, if $\mathbf{u}^k \in B(\mathbf{u}^*, \epsilon)$, the following relation is satisfied*

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\| \leq \hat{L}\|\mathbf{u}^k - \mathbf{u}^*\|^2.$$

Proof. Let us start by chosing $\epsilon = \min\{\delta, r\}$, so that both Lemma 4.1 and Assumption 4.4 holds. Since $\tilde{\mathcal{I}}^k = \mathcal{I}^k = \mathcal{I}^*$ and $\tilde{\mathcal{A}}^k = \mathcal{A}_0^k = \mathcal{A}^*$, for all $k > k_0$, and $\mathbf{u}^k \in B(\mathbf{u}^*, r)$, the GSDM system

$$\begin{pmatrix} \left(\nabla^2 f(\mathbf{u}^k)_{[\mathcal{I}^*,\mathcal{I}^*]} + \sigma\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}\right) & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \begin{pmatrix} \mathbf{w}^k_{\mathcal{I}^*} \\ \mathbf{w}^k_{\mathcal{A}_0^*} \end{pmatrix} = - \begin{pmatrix} \bar{\mathbf{v}}^k_{\mathcal{I}^*} \\ \mathbf{0}_{\mathcal{A}_0^*} \end{pmatrix},$$

is solvable for all $k > k_0$. Moreover, the updated iterate, with full step size, reads as

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{w}^k,$$

where

$$\begin{aligned} \mathbf{w}^k = \begin{pmatrix} \mathbf{w}^k_{\tilde{\mathcal{I}}^*} \\ \mathbf{w}^k_{\mathcal{A}_0^*} \end{pmatrix} &= \begin{pmatrix} \left(\nabla^2 f(\mathbf{u}^k)_{[\mathcal{I}^*,\mathcal{I}^*]} + \sigma\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}\right) & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix}^{-1} \begin{pmatrix} -\bar{\mathbf{v}}^k_{\mathcal{I}^*} \\ \mathbf{0}_{\mathcal{A}_0^*} \end{pmatrix} \\ &= \begin{pmatrix} -\left(H^k_{R_{[\mathcal{I}^*,\mathcal{I}^*]}}\right)^{-1} \bar{\mathbf{v}}^k_{\mathcal{I}^*} \\ \mathbf{0}_{\mathcal{A}_0^*} \end{pmatrix}. \end{aligned}$$

Therefore, given that $\mathbf{u}^k = \begin{pmatrix} \mathbf{u}^k_{\mathcal{I}^*} \\ \mathbf{u}^k_{\mathcal{A}_0^*} \end{pmatrix}$ and $\mathbf{u}^k_{\mathcal{A}_0^*} = \mathbf{0}_{\mathcal{A}_0^*}$, we get that

$$\mathbf{u}^{k+1} = \begin{pmatrix} \mathbf{u}^k_{\mathcal{I}^*} - \left(H^k_{R_{[\mathcal{I}^*,\mathcal{I}^*]}}\right)^{-1} \bar{\mathbf{v}}^k_{\mathcal{I}^*} \\ \mathbf{0}_{\mathcal{A}_0^*} \end{pmatrix}. \tag{4.71}$$

Let us analyze the difference $\|\mathbf{u}^{k+1} - \mathbf{u}^*\|$. From (4.71) and given that $\mathbf{u}^*_{\mathcal{A}_0^*} = \mathbf{0}_{\mathcal{A}_0^*}$, we

obtain that

$$\begin{aligned}\|\mathbf{u}^{k+1}-\mathbf{u}^*\| &= \left\|\mathbf{u}^k_{\tilde{\mathcal{I}}^*}-\left(H^k_{R_{[\tilde{\mathcal{I}}^*,\tilde{\mathcal{I}}^*]}}\right)^{-1}\bar{\mathbf{v}}^k_{\tilde{\mathcal{I}}^*}-\mathbf{u}^*_{\tilde{\mathcal{I}}^*}\right\| \\ &\leq \left\|\left(H^k_{R_{[\mathcal{I}^*,\mathcal{I}^*]}}\right)^{-1}\right\|\left\|H^k_{R_{[\mathcal{I}^*,\mathcal{I}^*]}}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})-\bar{\mathbf{v}}^k_{\mathcal{I}^*}\right\|.\end{aligned}$$

Thanks to Lemma 4.1 and since $-\bar{\mathbf{v}}^*_i=-\nabla_i f(\mathbf{u}^*)-\sigma\frac{\mathbf{u}^*_i}{\|\mathbf{u}^*_i\|}=0$, for all $i\in\mathcal{I}^*$, we then get that

$$\|\mathbf{u}^{k+1}-\mathbf{u}^*\| \leq C_H\left\|H^k_{R_{[\mathcal{I}^*,\mathcal{I}^*]}}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})-\bar{\mathbf{v}}^k_{\mathcal{I}^*}+\bar{\mathbf{v}}^*_{\mathcal{I}^*}\right\|. \tag{4.72}$$

Moreover, by replacing $H^k_{R_{[\mathcal{I}^*,\mathcal{I}^*]}}=\left(\nabla^2 f(\mathbf{u}^k)_{[\mathcal{I}^*,\mathcal{I}^*]}+\sigma\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}\right)$ in (4.72) we obtain that,

$$\|\mathbf{u}^{k+1}-\mathbf{u}^*\| \leq C_H\|\nabla^2 f(\mathbf{u}^k)_{[\mathcal{I}^*,\mathcal{I}^*]}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})+\sigma\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})-\bar{\mathbf{v}}^k_{\mathcal{I}^*}+\bar{\mathbf{v}}^*_{\mathcal{I}^*}\|. \tag{4.73}$$

We can also express the vector $-\bar{\mathbf{v}}^k_{\mathcal{I}^*}$ as follows:

$$-\bar{\mathbf{v}}^k_{\mathcal{I}^*}=-\nabla_{\mathcal{I}^*}f(\mathbf{u}^k)-\sigma Q(\mathbf{u}^k_{\mathcal{I}^*})\mathbf{u}^k_{\mathcal{I}^*}, \tag{4.74}$$

where $Q(\mathbf{u}^k_{\mathcal{I}^*})$ is a block-diagonal matrix, where each diagonal block is denoted by Q^k_i and given by:

$$Q^k_i=\frac{1}{\|\mathbf{u}^k_i\|}I_{n_i}, \qquad \forall i\in\mathcal{I}^*. \tag{4.75}$$

A related characterization follows for $\bar{\mathbf{v}}^*_{\mathcal{I}^*}$, i.e.,

$$\bar{\mathbf{v}}^*_{\mathcal{I}^*}=\nabla_{\mathcal{I}^*}f(\mathbf{u}^*)+\sigma Q(\mathbf{u}^*_{\mathcal{I}^*})\mathbf{u}^*_{\mathcal{I}^*}. \tag{4.76}$$

By plugging (4.74) and (4.76) in (4.73), and using the triangle inequality, we obtain:

$$\begin{aligned}\|\mathbf{u}^{k+1}-\mathbf{u}^*\| \leq C_H&\left\|\nabla^2 f(\mathbf{u}^k)_{[\mathcal{I}^*,\mathcal{I}^*]}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})-\nabla_{\mathcal{I}^*}f(\mathbf{u}^k)+\nabla_{\mathcal{I}^*}f(\mathbf{u}^*)\right\| \\ &+C_H\sigma\left\|\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})-Q(\mathbf{u}^k_{\mathcal{I}^*})\mathbf{u}^k_{\mathcal{I}^*}+Q(\mathbf{u}^*_{\mathcal{I}^*})\mathbf{u}^*_{\mathcal{I}^*}\right\|.\end{aligned} \tag{4.77}$$

We analyze the first term of the right-hand side of (4.77). Since $\nabla^2 f$ is locally Lipschitz continuous at $\mathbf{u}^*$ (with Lispchitz constant L_f), by using the integral form of the Mean Value Theorem, we get:

$$\left\|\nabla^2 f(\mathbf{u}^k)_{[\mathcal{I}^*,\mathcal{I}^*]}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})-\nabla_{\mathcal{I}^*}f(\mathbf{u}^k)+\nabla_{\mathcal{I}^*}f(\mathbf{u}^*)\right\| \leq \frac{L_f}{2}\|\mathbf{u}^k-\mathbf{u}^*\|^2. \tag{4.78}$$

For the second term in (4.77) we proceed as follows. By the equivalence of norms property $\|\cdot\|\leq\|\cdot\|_{1,2}$, and based on the definitions of $\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}$ in (4.70) and Q in (4.75),

it follows that:

$$\left\|\Gamma^k_{[\mathcal{I}^*,\mathcal{I}^*]}(\mathbf{u}^k_{\mathcal{I}^*}-\mathbf{u}^*_{\mathcal{I}^*})-Q(\mathbf{u}^k_{\mathcal{I}^*})\mathbf{u}^k_{\mathcal{I}^*}+Q(\mathbf{u}^*_{\mathcal{I}^*})\mathbf{u}^*_{\mathcal{I}^*}\right\| \\ \leq \sum_{i\in\mathcal{I}^*}\left\|\left(\frac{1}{\|\mathbf{u}^k_i\|}I_{n_i}-\frac{\mathbf{u}^k_i\mathbf{u}^{k\top}_i}{\|\mathbf{u}^k_i\|^3}\right)(\mathbf{u}^k_i-\mathbf{u}^*_i)-\frac{\mathbf{u}^k_i}{\|\mathbf{u}^k_i\|}+\frac{\mathbf{u}^*_i}{\|\mathbf{u}^*_i\|}\right\|. \quad (4.79)$$

Given a vector $\mathbf{z}\neq\mathbf{0}$, we have that the function $G:\mathbf{z}\to\frac{\mathbf{z}}{\|\mathbf{z}\|}$ from $\mathbb{R}^l\backslash\{\mathbf{0}\}$ to itself is differentiable with Fréchet derivative G' given by $G'(\mathbf{z})=\frac{1}{\|\mathbf{z}\|}\left(I-\frac{\mathbf{z}\mathbf{z}^\top}{\|\mathbf{z}\|^2}\right)$. Moreover, $\mathbf{z}\to G'(\mathbf{z})$ is locally Lipschitz continuous from $\mathbb{R}^l\backslash\{\mathbf{0}\}$ to $\mathbb{R}^{l\times l}\backslash\{\mathbf{0}\}$ (see[66]). Therefore, we can rewrite (4.79) by means of function G and its derivative G':

$$\left\|\Gamma^k_{[\tilde{\mathcal{I}}^*,\tilde{\mathcal{I}}^*]}(\mathbf{u}^k_{\tilde{\mathcal{I}}^*}-\mathbf{u}^*_{\tilde{\mathcal{I}}^*})-Q(\mathbf{u}^k_{\tilde{\mathcal{I}}^*})\mathbf{u}^k_{\tilde{\mathcal{I}}^*}+Q(\mathbf{u}^*_{\tilde{\mathcal{I}}^*})\mathbf{u}^*_{\tilde{\mathcal{I}}^*}\right\| \leq \sum_{i\in\tilde{\mathcal{I}}^*}\left\|G'(\mathbf{u}^k_i)(\mathbf{u}^k_i-\mathbf{u}^*_{\tilde{\mathcal{I}}^*})-G(\mathbf{u}^k_i)+G(\mathbf{u}^*_i)\right\|. \quad (4.80)$$

Following the same procedure as for term (4.78), we obtain that

$$\begin{aligned}\left\|\Gamma^k_{[\tilde{\mathcal{I}}^*,\tilde{\mathcal{I}}^*]}(\mathbf{u}^k_{\tilde{\mathcal{I}}^*}-\mathbf{u}^*_{\tilde{\mathcal{I}}^*})-Q(\mathbf{u}^k_{\tilde{\mathcal{I}}^*})\mathbf{u}^k_{\tilde{\mathcal{I}}^*}\right. &\left.+Q(\mathbf{u}^*_{\tilde{\mathcal{I}}^*})\mathbf{u}^*_{\tilde{\mathcal{I}}^*}\right\| \\ &\leq \sum_{i\in\tilde{\mathcal{I}}^*}\frac{L_2}{2}\|\mathbf{u}^k_i-\mathbf{u}^*_i\|^2 \\ &\leq M\frac{L_G}{2}\|\mathbf{u}^k-\mathbf{u}^*\|^2,\end{aligned} \quad (4.81)$$

where L_G is the Lipschitz constant of G' and $M:=|\mathcal{I}^*|$. Plugging (4.78) and (4.81) into (4.77) yields

$$\|\mathbf{u}^{k+1}-\mathbf{u}^*\|\leq\frac{C_H}{2}(L_f+M\sigma L_G)\|\mathbf{u}^k-\mathbf{u}^*\|^2=\hat{L}\|\mathbf{u}^k-\mathbf{u}^*\|^2,$$

where $\hat{L}=\frac{C_H}{2}(L_f+M\sigma L_G)$. □

Theorem 4.9. *Let Assumption 4.4 hold. Then, there exists $\hat{\epsilon}>0$ such that if any $\mathbf{u}^0\in B(\mathbf{u}^*,\hat{\epsilon})$, the sequence generated by the GSDM method, with full step size, converges to $\mathbf{u}^*$ with Q-quadratic converge.*

Proof. The proof follows by induction. Let $\hat{\epsilon}$ be small enough so that the conclusions of Theorem 4.8 hold. Then, if $\mathbf{u}^0\in B(\mathbf{u}^*,\hat{\epsilon})$ we have that

$$\|\mathbf{u}^1-\mathbf{u}^*\|\leq\hat{L}\|\mathbf{u}^0-\mathbf{u}^*\|^2. \quad (4.82)$$

We further reduce the radius $\hat{\epsilon}$ such that $\|\mathbf{u}^0-\mathbf{u}^*\|\leq\min\{\epsilon,\frac{1}{2\hat{L}}\}$. Thus, from (4.82), we have

$$\|\mathbf{u}^1-\mathbf{u}^*\|\leq\hat{L}\|\mathbf{u}^0-\mathbf{u}^*\|^2\leq\hat{L}\frac{1}{2\hat{L}}\|\mathbf{u}^0-\mathbf{u}^*\|=\frac{1}{2}\|\mathbf{u}^0-\mathbf{u}^*\|. \quad (4.83)$$

Hence, if $\hat{\epsilon}=\min\{\epsilon,\frac{1}{2\hat{L}}\}$, then $\mathbf{u}^0\in B(\mathbf{u}^*,\epsilon)$ implies $\mathbf{u}^1\in B(\mathbf{u}^*,\epsilon)$.

By induction we then obtain that if $\mathbf{u}^k \in B(\mathbf{u}^*, \epsilon)$, then

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\| \leq \frac{1}{2}\|\mathbf{u}^k - \mathbf{u}^*\| \leq \left(\frac{1}{2}\right)^{k+1} \|\mathbf{u}^0 - \mathbf{u}^*\|. \tag{4.84}$$

Consequently, the sequence $\{\mathbf{u}^k\}$ converges to $\mathbf{u}^*$. Additionally, from Theorem 4.8, we obtain that

$$\|\mathbf{u}^{k+1} - \mathbf{u}^*\| \leq \hat{L}\|\mathbf{u}^k - \mathbf{u}^*\|^2. \tag{4.85}$$

which implies the Q-quadratic convergence of the GSDM method. □

4.6 Numerical implementation and computational experiments

Let us comment on several aspects of the numerical implementation of GSDM Algorithm and its associated parameters.

Line search strategy

In our examples, we employ a generalized Armijo condition designed for locally Lipschitz continuous functions which are not necessarily convex (see [121]). In our context, the generalized Armijo decrease condition for ψ is given by:

$$\psi(\mathbf{u}^{k+1}) \leq \psi(\mathbf{u}^k) + c_0 s^k \psi'(\mathbf{u}^k, \mathbf{w}^k). \tag{4.86}$$

where s^k is the line-search step and c_0 is a positive fixed constant. Inequality (4.86) extends the traditional Armijo condition through a reformulation using the directional derivative. In the experiments below, the backtracking scheme utilizes the constant value $c_0 = 1e-4$, the initialization $s^k = 1$ and the upsate $s^k \leftarrow \min\left(\frac{s^k}{d}, 1\right) d^j$ for $j = 0, 1, \ldots$, until the sufficient decrease (4.86) is satisfied. In our case we set $d = 0.8$.

Stopping criteria

Algorithm 5 terminates once the necessary optimality condition (4.32) is satisfied and the directional derivative at $\mathbf{u}^k$ along $\mathbf{w}^k$ falls below a prescribed tolerance, indicating that $\mathbf{w}^k$ is no longer a descent direction. Specifically, based on the necessary optimality

conditions (4.32), we check that:

$$sc_1 := \max_{\{i:\|\mathbf{u}_i^k\|>0\}} \left\| \nabla_i f(\mathbf{u}^k) + \sigma \frac{\mathbf{u}_i^k}{\|\mathbf{u}_i^k\|} \right\| < \text{tol}_1, \tag{SC1}$$

$$sc_2 := \max_{\{i:\|\mathbf{u}_i^k\|=0\}} \|\nabla_i f(\mathbf{u}^k)\| - \sigma \leq \text{tol}_1, \tag{SC2}$$

and for the descent direction $\mathbf{w}^k$ we check that:

$$\psi'(\mathbf{u}^k, \mathbf{w}^k) > -\text{tol}_2. \tag{SC3}$$

In our experiments, the GSDM algorithm terminates once the verification of (SC1), (SC2) and (SC3) holds true for $\text{tol}_1 = 1e-4$ and $\text{tol}_2 = 1e-7$.

Sparsity

The computation of the active index-set $\mathcal{A}_0^k$ in Algorithm 5 is conducted by checking approximately the conditions for activity, i.e., if $\|\mathbf{u}_i^k\| \leq \epsilon$ and $\|\nabla_i f(\mathbf{u}^k)\| - \sigma \leq \text{tol}_1$, then index i is included in the active index-set $\mathcal{A}_0^k$.

4.6.1 Nonconvex Support Vector Machine

We consider the following support vector machine problem:

$$\min_{\mathbf{u}\in\mathbb{R}^m} \psi(\mathbf{u}) := \frac{1}{l}\sum_{j=1}^{l} \Big(1 - \tanh(b_j\langle \mathbf{a}_j, \mathbf{u}\rangle)\Big) + \sigma\|\mathbf{u}\|_{1,2}. \tag{SVM}$$

Let us recall that $\mathbf{u} \in \mathbb{R}^m$ is a vector composed of p predetermined groups $\mathbf{u}_i$ such that each $\mathbf{u}_i \in \mathbb{R}^{n_i}$ and $\sum_{i=1}^p n_i = m$. Moreover, $b_j \in \{-1, 1\}$ are labels and $\mathbf{a}_j \in \mathbb{R}^m$ are data points containing the information of the feature vectors for $j = 1, \cdots, l$.

Datasets and Grouping

To evaluate the performance of GSDM on problem (SVM), we consider three classification examples: two benchmark datasets—*CDC Diabetes Health Indicators* and *Wine Quality*—sourced from the Kaggle [24] and UCI [89] repositories, and a synthetic binary classification dataset generated using the Scikit-learn library in Python.

The *CDC Diabetes Health Indicators* benchmark [24] is a binary classification data set containing lab test results and a health-related telephone survey. It contains 70692 instances and 21 features. The dataset is evenly divided in two categories with 50% of respondents having no diabetes and the other 50% diagnosed with either predia-

betes or diabetes, i.e., the target variable for classification is whether a patient has diabetes/pre-diabetes or not. The 21 features consist in demographics (race, sex), personal information (income, educations) and health history (drinking, smoking, mental health, physical health, etc.).

The *Wine Quality* dataset [35] contains the results of physicochemical tests (such as volatile acidity, pH, sulphates, and chlorides) used to evaluate the quality of wines. The target variable represents wine quality on a scale from 0 to 10. The dataset comprises 4,898 instances and 11 input features. To adapt this dataset for binary classification, scores greater than 5 are labeled as class 1, while scores of 5 or lower are labeled as -1.

Scikit-learn synthetic classification dataset is designed with a controlled structure to resemble scenarios for classification problems. We designed the dataset for binary classification such that it contains 1000 samples and 200 features. From the 200 features only 50 features are informative. An additional 20 features are redundant, i.e., linear combinations of the informative ones. The remaining 130 features are noise. This way, the model has to identify and rely only on a small subset of truly useful features, promoting sparsity.

Grouping. In order to group the datasets features we use the Spectral Clustering algorithm to identify groups of features that are closely related. This strategy captures non-linear relationships (see [93] and [118]). We applied the Spectral Clustering algorithm, implemented in the Scikit-learn machine learning library in Python, to the features of the datasets in order to identify groups of strongly correlated variables. First, the Pearson correlation coefficients between all pairs of features were computed to form a similarity matrix, where absolute correlation values were used. This matrix was then provided as input to the Spectral Clustering algorithm, which grouped the features into distinct clusters based on their correlation structure (see Table 4.2).

Analysis of GSDM Parameters: ϵ and γ.

To examine GSDM internal parameters, we evaluate its behavior over several values for the Huber regularization parameter γ and the global convergence parameter ϵ, under four fixed values of the penalization parameter σ. The results, presented in Table 4.3 show that the algorithm is stable and convergent across all tested configurations, the descent direction $\mathbf{w}^k$ satisfies $\psi'(\mathbf{u}^k, \mathbf{w}^k) \approx -10^{-8}$ or lower. Notably, larger values of σ lead to higher sparsity levels, with $\sigma =$ 1e-1 consistently yielding 80% sparsity in only 9 iterations, regardless of (ϵ, γ). However, for $\sigma =$ 5e-4, larger values of ϵ (e.g. 1e-3) lead to fewer iterations, although the resulting cost is slightly higher compared to the values obtained with smaller ϵ. Overall, the method shows consistency with respect to the inherent parameters ϵ and γ.

CDC Diabetes Health Indicators dataset		
Group	No. of features	Features
1	2	have any healthcare coverage, couldn't afford a doctor in a year
2	3	high blood pressure, high cholesterol, age
3	6	BMI, physical activity, general health, mental health, physical health, have difficulty walking
4	1	smoker
5	1	heavy alcohol consumption
6	1	sex
7	2	fruits consumption, veggies consumption
8	2	stroke, heart disease
9	1	no cholesterol check
10	2	education, income
Wine Quality dataset		
Group	No. of features	Features
1	1	alcohol
2	2	residual sugar, density
3	2	free sulfur dioxide, total sulfur dioxide
4	4	fixed acidity, volatile acidity, citric acid, pH
5	2	chlorides, sulphates
Scikit-learn synthetic dataset		
No. groups	No. of features	Features
50	1-16	each of the 50 groups contains between 1 and 16 synthetic features

Table 4.2: Spectral clustering of features across three datasets.

Index-sets Identification Property of GSDM for problem (SVM)

With the two remaining datasets, *Wine Quality* and the *Scikit-learn synthetic dataset*, we evaluate the performance of the GSDM algorithm by analyzing the index-sets identification.

Figures 4.4 illustrate the evolution of the index-set sizes produced by GSDM for the *Wine Quality* and *Scikit-learn synthetic* datasets, respectively. The x-axis represents the iteration number, while the y-axis shows the size of the index-sets $\mathcal{A}_{\nabla}^{k}$, $\mathcal{A}_{0}^{k}$, $\mathcal{J}^{k}$, and $\tilde{\mathcal{I}}^{k}$. The figure shows each index-set as a separate bar next to each other for every iteration. GSDM quickly identifies the active and inactive groups corresponding to the optimal solution, ultimately leading to an empty $\mathcal{A}_{\nabla}^{k}$, indicating that the sparse groups have been correctly classified. Moreover, toward the final iterations, the index set $\mathcal{J}^{k}$—which is associated with the active-set prediction phase—becomes empty, consistent with the theoretical result provided by Theorem 4.7. These results confirm that, from a certain iteration onward, the algorithm successfully achieves $\tilde{\mathcal{I}}^{k} = \mathcal{I}^{k}$ and $\tilde{\mathcal{A}}^{k} = \mathcal{A}_{0}^{k} = \mathcal{A}^{k}$.

ϵ	γ	σ = 5e-4				σ = 1e-3				σ = 1e-2				σ = 1e-1			
		Cost	ψ'	It.	Sparsity (%)	Cost	ψ'	It.	Sparsity (%)	Cost	ψ'	It.	Sparsity (%)	Cost	ψ'	It.	Sparsity (%)
1e-3	1e3	0.51127	-9.80e-08	11	10	0.51603	-6.09e-09	10	20	0.56441	-4.03e-12	10	30	0.73126	-7.42e-14	9	80
	1e4	0.51127	-3.12e-08	10	10	0.51603	-5.35e-08	9	20	0.56441	-7.72e-08	9	30	0.73126	-7.42e-14	9	80
	1e5	0.51127	-3.12e-08	10	10	0.51603	-5.35e-08	9	20	0.56441	-7.72e-08	9	30	0.73126	-7.42e-14	9	80
1e-6	1e3	0.51053	-2.59e-09	22	10	0.51603	-7.27e-08	11	20	0.56441	-3.15e-12	12	30	0.73126	-7.42e-14	9	80
	1e4	0.51053	1.78e-10	43	10	0.51603	-7.27e-08	11	20	0.56441	-1.04e-09	11	30	0.73126	-7.42e-14	9	80
	1e5	0.51053	-1.79e-10	43	10	0.51603	-7.27e-08	11	20	0.56441	-1.04e-09	11	30	0.73126	-7.42e-14	9	80
1e-8	1e3	0.51053	-9.30e-08	25	10	0.51603	-7.48e-08	21	20	0.56441	-3.39e-09	12	30	0.73126	-7.42e-14	9	80
	1e4	0.51053	-2.66e-12	46	10	0.51603	-7.48e-08	21	20	0.56441	-1.04e-09	11	30	0.73126	-7.42e-14	9	80
	1e5	0.51053	-2.66e-12	46	10	0.51603	-7.48e-08	21	20	0.56441	-1.04e-09	11	30	0.73126	-7.42e-14	9	80
0	1e3	0.51053	-9.19e-08	28	10	0.51603	-8.07e-08	21	20	0.56441	-3.39e-09	12	30	0.73126	-7.42e-14	9	80
	1e4	0.51053	-1.35e-08	46	10	0.51603	-8.07e-08	21	20	0.56441	-1.04e-09	11	30	0.73126	-7.42e-14	9	80
	1e5	0.51053	-1.35e-08	46	10	0.51603	-8.07e-08	21	20	0.56441	-1.04e-09	11	30	0.73126	-7.42e-14	9	80

Table 4.3: Comparison of GSDM performance on *CDC Diabetes Health Indicators* for varying (ϵ, γ) parameters and σ = 5e-4, 1e-3, 1e-2, and 1e-1. $\psi' = \psi'(\mathbf{u}^k, \mathbf{w}^k)$.

Figure 4.5 illustrates the convergence behavior of the GSDM method for different values of the regularization parameter σ, applied to the *Wine Quality* and *Scikit-learn synthetic* datasets. In both cases, the absolute values of the directional derivative at $\mathbf{u}^k$ along $\mathbf{w}^k$ (Figure 4.5) show a consistent decrease, demonstrating that the method satisfies criterion SC3.

4.6.2 Semilinear Elliptic Optimal Control problems

Consider the elliptic optimal control problem (OCP):

$$\min_{u \in L^2(\Omega)} \psi := \frac{1}{2}\|S(u) - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} + \sigma\|u\|_{1,2}, \tag{OCP}$$

where the state variable, given by $y = S(u) \in H^1_0(\Omega)$, is the solution of the semilinear elliptic problem

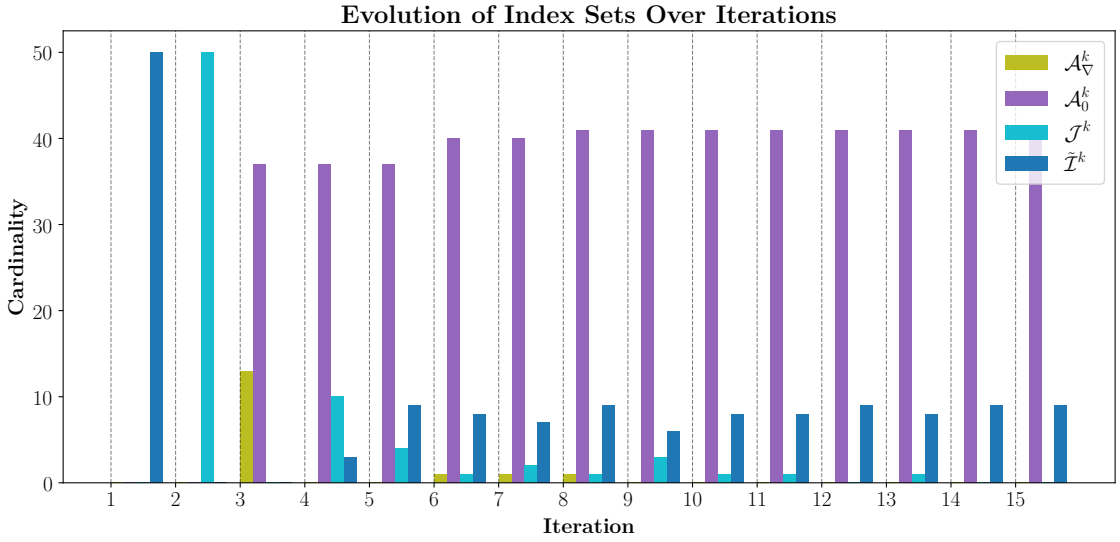
$$-\Delta y + y^3 = u \ \text{ in } \Omega, \ \ u = 0 \text{ on } \Gamma. \tag{4.88}$$

Given the unit square domain $\Omega = (0,1) \times (0,1)$, the sparsity patterns promoted by the $\|\cdot\|_{1,2}$ norm in this problem are defined over the vertical cross-sections of Ω. Consequently, each group corresponds to the restriction of u to these cross-sections of the unit square.

Problem (OCP) requires a discretization process to fit our formulation. We discretize the domain $\Omega = (0,1) \times (0,1)$ using a 64×64 grid with mesh size step $h = 1/(64+1)$. We use the five-point stencil to write the finite difference approximation and a trapezoidal quadrature rule to compute the integrals. This process leads us to a problem of the form

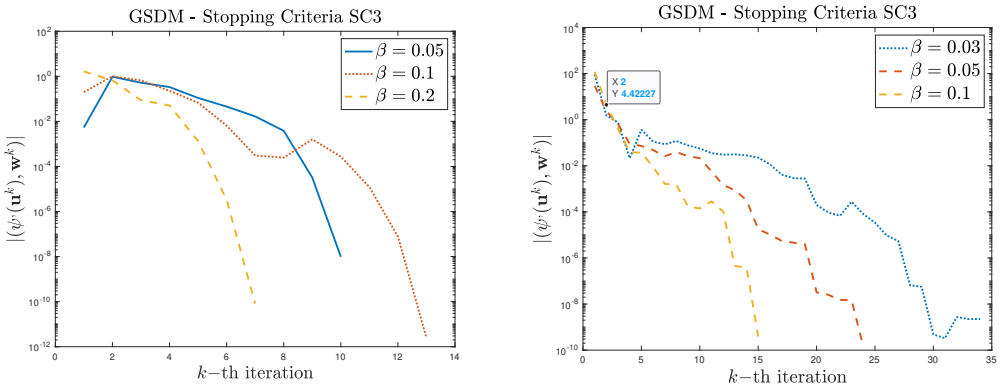
$$\min_{\mathbf{u} \in \mathbb{R}^m} \frac{1}{2}(\mathbf{S}(\mathbf{u}) - \mathbf{y}_d)^\top M_1 (\mathbf{S}(\mathbf{u}) - \mathbf{y}_d) + \frac{\alpha}{2}\mathbf{u}^\top M_1 \mathbf{u} + \sigma \sum_{i=1}^{p} h\sqrt{\mathbf{u}_i^\top M_2 \mathbf{u}_i},$$

(a) *Wine Quality* dataset – group index-set classification: Each index-set is a separate bar for every iteration. Penalization parameter $\sigma = 2\text{e}{-1}$.



(b) *Scikit-learn synthetic* dataset – group index-set classification: Each index-set is a separate bar for every iteration. Penalization parameter $\sigma = 1\text{e}{-1}$.

Figure 4.4: Group index-set classification over iterations for two datasets. GSDM's parameters: $\gamma = 1\text{e}5$, $\epsilon = 1\text{e}{-8}$,



(a) *Wine Quality* (b) *Scikit-learn synthetic classification*

Figure 4.5: GSDM performace in (SVM) problem: history of absolute values of the directional derivative, $|\psi'(\mathbf{u}^k, \mathbf{w}^k)|$ (SC3).

where the matrices M_1 and M_2 are positive definite matrices resulting from numerical integration and $\mathbf{S}$ corresponds to the (discrete) solution nonlinear mapping that solves equation (4.88) numerically for a given discrete control $\mathbf{u}$.

For this experiment we use the desired state $\mathbf{y}_d(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2)\exp(2x_1)/6$ and the penalization parameters $\alpha = 1e-6$ and $\sigma = 1e-4$.

Analysis of Second-Order Information for problem (OCP)

A key feature of GSDM is the use of a reduced second-order matrix to accelerate the descent direction within the inactive group set $\tilde{\mathcal{I}}^k$. The reduced matrix is defined as $H^k_{R_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}} := B^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]} + \sigma\Gamma^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$, where $B^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$ corresponds to the exact reduced Hessian or an approximation matrix of it and Γ^k arises from the Huber regularization of the $\|\cdot\|_{1,2}$ norm. In this experiment, we assess the performance of GSDM using three different constructions of the second-order matrix. In the first case, B^k is taken as the exact reduced Hessian, i.e., $B^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]} = \nabla^2 f(\mathbf{u}^k)_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$ in addition to the regularization matrix Γ^k. The second alternative considers only the exact Hessian without regularization, that is, $H^k_{R_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}} := \nabla^2 f(\mathbf{u}^k)_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$. The third option uses a BFGS approximation matrix B^k in place of the true Hessian, again combined with the regularization term. The last experiment performs a run of the algorithm with no second-order information. This comparison (see Table 4.4) enables an evaluation of the impact of second-order accuracy and regularization on the overall performance of the algorithm. This analysis is presented in conjunction with the influence of different initial guesses. Across all initializations, using the full second-order matrix $\nabla^2 f(\mathbf{u}^k)_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]} + \sigma\Gamma^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$ consistently leads to faster convergence, with significantly fewer iterations and lower computational time compared to the alternative approximation of the BFGS matrix. The algorithm's performance using only the reduced Hessian matrix—excluding the regularization com-

ponent $\Gamma^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$- results in a substantial increase in the number of iterations, exceeding a fivefold rise. The final experiment was terminated after 2550 iterations without meeting the convergence criteria defined in (SC1), (SC2) and (SC3), resulting in a cost value of 0.02000 and an absolute error of 0.01788 relative to the solution obtained using the full reduced second-order matrix.

Regarding the initial point, the null and the Poisson-based initial guesses yield the most efficient performance, converging in 15 and 20 iterations, respectively. In contrast, the random initialization and the constant initialization $u^0(x_1, x_2) = 1$ require substantially more iterations—up to 432 in the worst case—when paired with the approximate BFGS matrix.

These results highlight the importance of incorporating accurate second-order information and a well-informed initial guess to enhance the efficiency and robustness of GSDM.

Initial guess	Matrix $H^k_{R_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}}$		GSDM performance				
	Smooth part $B^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$	Nonsmooth $\Gamma^k_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$	Cost	$\psi'(\mathbf{u}^k, \mathbf{w}^k)$	It.	Sparsity (%)	Time (s)
Poisson solution $-\Delta u^0 = y_d$	reduced Hessian	yes	0.00212	-1.12e-14	20	32.8	63
	reduced Hessian	no	0.00212	-3.15e-14	245	31.2	770
	BFGS	yes	0.00215	-3.21e-08	80	37.5	51
	no	no	0.02000	-5.28e-06	2550	3.1	-
$u^0(x_1, x_2) = 0$	reduced Hessian	yes	0.00212	-5.70e-08	15	32.8	48
	reduced Hessian	no	0.00212	-2.69e-12	238	34.4	771
	BFGS	yes	0.00216	-7.79e-08	80	34.4	63
	no	no	0.02000	-5.28e-06	2550	3.1	-
$u^0(x_1, x_2) = 1$	reduced Hessian	yes	0.00212	-9.08e-13	32	34.4	122
	reduced Hessian	no	0.00212	-4.27e-13	233	31.2	721
	BFGS	yes	0.00217	-7.64e-08	432	40.6	303
	no	no	0.02001	-5.28e-06	2550	3.1	-
$u^0(x_1, x_2) = rand$	reduced Hessian	yes	0.00213	-2.87e-9	38	40.6	136
	reduced Hessian	no	0.00213	-7.39e-15	137	39.0	448
	BFGS	yes	0.00216	-8.02e-08	235	35.9	158
	no	no	0.02001	-5.28e-06	2550	3.1	-

Table 4.4: Second order reduced matrix $H^k_{R_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}}$ configurations and initial guess analysis in (OCP) with parameters: $\sigma = 1e{-}4$ and $\alpha = 1e{-}6$. GSDM's parameters $\gamma = 1e5$ and $\epsilon = 1e{-}8$

Step size and Index-sets Identification relation for problem (OCP)

By initializing the algorithm with the solution of the Poisson problem, using the matrix $\nabla^2 f(\mathbf{u})_{[\tilde{\mathcal{I}}^k,\tilde{\mathcal{I}}^k]}$, and incorporating the full second-order information $\nabla^2 f(\mathbf{u}^k) + \sigma\Gamma^k$, Table 4.5 presents the evolution of the index sets $\mathcal{A}_\nabla^k$, $\mathcal{A}^k$, $\mathcal{J}^k$, and $\tilde{\mathcal{I}}^k$ across selected iterations of a single GSDM run. The table reports the number of groups in each index set at iteration k, alongside the step size computed via backtracking, shown in the final column. In the final stages of optimization (e.g., iteration 19), the algorithm achieves sufficient decrease in the objective function using a unit step size, and the active-set $\mathcal{J}^k$ becomes empty, in agreement with the convergence behavior described in Theorem 5.2. In addition, the last column of Table 4.5 shows the number of iterations performed by the line search procedure.

The column labeled "#System" represents the size of the reduced second-order system solved at each iteration. As the number of elements in $\mathcal{A}_0^k$ increases, the dimension of the system decreases. This reflects a reduction in the number of unknowns, and highlights the efficiency gains of GSDM as the algorithm progresses.

Active-set prediction phase Analysis for problem (OCP)

We also tested the GSDM algorithm without the active-set projection phase. The algorithm was initialized with the Poisson solution and parameters: $\sigma = 1e-4$ and $\alpha = 1e-6$. The results indicate that, despite progressing through several iterations, the method struggles to fully recover the correct sparsity pattern. Specifically, the set $\tilde{\mathcal{I}}^k$, which corresponds to the inactive groups, remains larger than expected, failing to match the true number of zero groups in the solution. At iteration 500, for example, only three groups have been identified as active via $\mathcal{A}_0^k$, with no entries detected in the active-set prediction set $\mathcal{J}^k$, and with the directional derivative $\psi'(\mathbf{u}^k, \mathbf{w}^k)$ still away from zero. Although the cost function value steadily decreases, reaching 0.00213 at iteration 500, the lack of full identification of active sets highlights the importance of the prediction phase in accelerating convergence and promoting accurate sparsity recovery. The algorithm was stopped after 600 iterations and failed to identify the total number of sparse groups and achieve the stopping criteria. In contrast, GSDM with the active-set projection phase identified 21 sparse groups after 20 iterations (see Table 4.5).

Comparison with the Semi-Smooth Newton method- SSN and convergence behaviour.

We compare the GSDM algorithms with the SSN method proposed in [64]. This SSN method is a damped version tailored for group sparse optimal control problems.

GSDM applied to (OCP)									
It. k	$\mathcal{A}_{\nabla}^k$	$\mathcal{A}_0^k$	$\mathcal{J}^k$	$\tilde{\mathcal{I}}^k$	Cost	$\psi'(\mathbf{u}^k, \mathbf{w}^k)$	#System	s^k	Line search it.
1	0.0	0.0	0.0	64.0	0.04223	-8.11e-02	4096	1.00	1
2	0.0	0.0	16.0	48.0	0.00336	-1.73e-03	3072	0.80	2
3	0.0	0.0	37.0	27.0	0.00329	-1.06e-02	1728	0.09	6
4	0.0	0.0	23.0	41.0	0.00279	-2.08e-03	2624	0.21	5
5	0.0	0.0	20.0	44.0	0.00263	-1.03e-03	2816	1.00	1
6	12.0	8.0	6.0	38.0	0.00255	-5.90e-04	2432	1.00	1
7	8.0	6.0	23.0	27.0	0.00235	-4.74e-03	1728	0.09	6
8	1.0	5.0	9.0	49.0	0.00231	-3.75e-04	3136	1.00	1
9	0.0	14.0	28.0	22.0	0.00225	-4.23e-04	1408	0.03	7
10	0.0	14.0	4.0	46.0	0.00225	-1.44e-04	2944	1.00	1
11	5.0	13.0	24.0	22.0	0.00216	-1.05e-03	1408	0.09	6
12	0.0	13.0	17.0	34.0	0.00214	-2.56e-04	2176	0.01	8
13	0.0	13.0	7.0	44.0	0.00214	-4.35e-05	2816	0.41	4
14	0.0	13.0	7.0	44.0	0.00214	-8.39e-05	2816	0.09	6
15	0.0	13.0	3.0	48.0	0.00214	-3.18e-05	3072	1.00	1
16	0.0	16.0	5.0	43.0	0.00213	-1.30e-05	2752	1.00	1
17	0.0	21.0	13.0	30.0	0.00212	-1.03e-04	1920	0.01	8
18	0.0	21.0	6.0	37.0	0.00212	-2.93e-05	2368	0.03	7
19	0.0	21.0	**0.0**	43.0	0.00212	-2.36e-06	2752	**1.00**	1
20	0.0	21.0	**0.0**	43.0	0.00212	-1.12e-14	2752	**1.00**	1

Table 4.5: Single run of GSDM applied to (OCP). The algorithm was initialized with the Poisson solution. Model's parameters: $\sigma = 1e-4$ and $\alpha = 1e-6$. GSDM's parameters $\gamma =$ 1e5 and $\epsilon =$ 1e−8

Initial guess $u^0(x_1, x_2) =$	Cost		Sparsity (%)		Iterations		Time	
	GSDM	SSN	GSDM	SSN	GSDM	SSN	GSDM	SSN
0	2.1265×10^{-3}	2.1215×10^{-3}	34.3	31.3	25	14	78.6	88.3
20	2.1222×10^{-3}	2.1215×10^{-3}	35.9	31.3	40	18	129.7	116.9
80	2.1240×10^{-3}	2.1215×10^{-3}	34.3	31.3	42	20	138.5	132.4
200	2.1235×10^{-3}	x	35.9	x	33	x	104.50	x

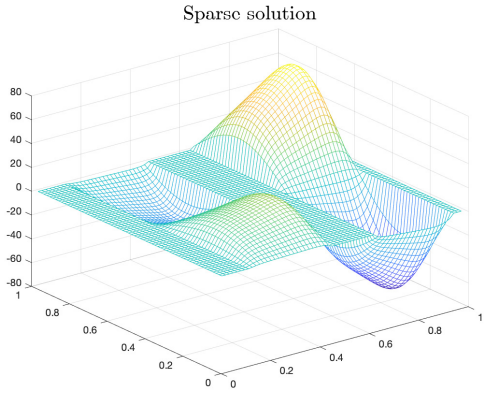
Table 4.6: Comparison of GSDM and SSN methods applied to solve (OCP) with a semi-linear elliptic PDE. Model's parameters: $\sigma = 1e-4$ and $\alpha = 1e-6$. GSDM's parameters $\gamma = 1e5$ and $\epsilon = 1e{-}7$. The symbol x indicates that the algorithm failed to converge

We compare both methods by solving the semi-linear elliptic optimal control problem (OCP) subject to (4.88).

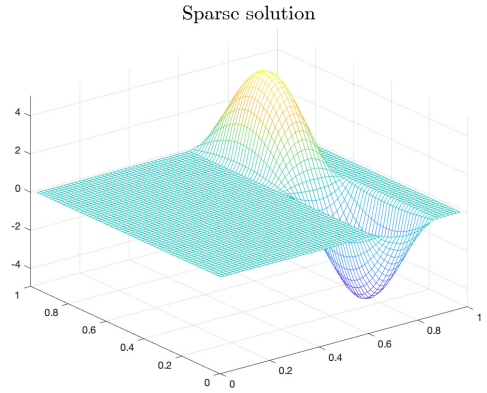
First, it is important to note that the SSN method solves a system twice the size of the one solved by GSDM. When discretizing the domain Ω with a 64×64 grid, GSDM handles a system of 4096 variables, whereas the SSN system consist of 8192 variables.

In Table 4.6, we compare the both algorithm performances using various initial guesses. The results demonstrate that GSDM exhibits robust convergence behavior even when starting far from the solution. As the initial guess moves further away (e.g., $u^0(x_1, x_2) = 200$), GSDM maintains convergence, requiring a moderate number of iterations and preserving the sparsity structure of the solution. The SSN method successfully converges for closer initializations but fails to converge when the initial guess is very distant from the optimal solution (e.g., $u^0(x_1, x_2) = 200$). This case is marked with an x symbol. These experiments highlight the advantage of GSDM's globalization strategy when the initial solution is far from the optimal state.

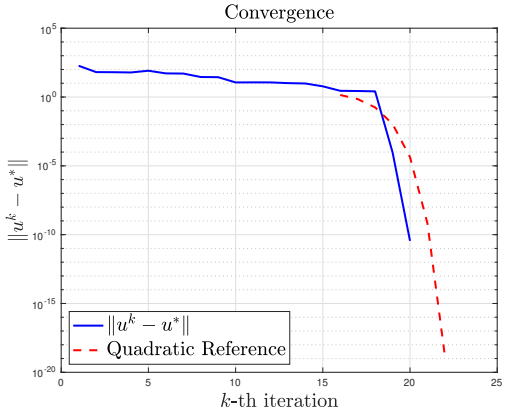
In addition, we perform a final test to illustrate the convergence behavior of the GSDM algorithm. A reference solution obtained after 200 iterations was used as the ground truth and the algorithm was initialized with the Poisson solution. In Figure 4.6, (c)-(d), we illustrate the convergence behavior of GSDM under two different parameter settings for α and σ. The error $\|u^k - u^*\|$ plotted on a logarithmic scale, decreases steadily over the iterations. Eventually, a sharp drop is observed, indicating that once the algorithm enters a neighborhood of the solution, it transitions to a superlinear or nearly quadratic rate of convergence. Finally, in Figure 4.6, (a)-(b), the optimal control solution obtained with GSDM is displayed for the different parameters α and σ.



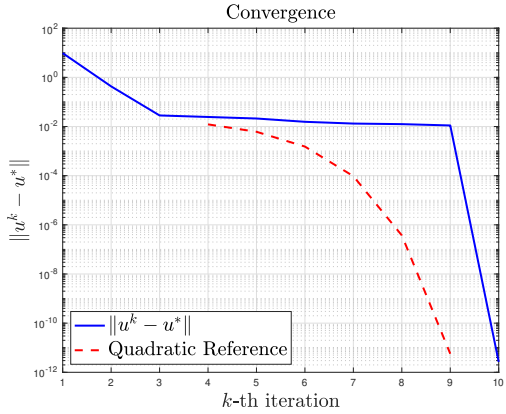
(a) Sparse control. Parameters: $\sigma = 1e{-}4$, $\alpha = 1e{-}6$.



(b) Sparse control. Parameters: $\sigma = 5e{-}3$, $\alpha = 5e{-}4$.



(c) Convergence. Parameters: $\sigma = 1e{-}4$, $\alpha = 1e{-}6$.



(d) Convergence. Parameters: $\sigma = 5e{-}3$, $\alpha = 5e{-}4$.

Figure 4.6: Semi-linear optimal control problem solved with GSDM: (a)-(b) Sparse optimal controls and (c)-(d) convergence behavior for different penalization parameters. GSDM's parameters: $\gamma = 1e5$ and $\epsilon = 1e{-}8$.

Chapter 5

Conclusions

In this thesis, we analyzed and designed second-order algorithms with applications to the exact penalization of the incompressibility condition in the Bingham flow problem and to group-sparse penalized problems. The convergence of both algorithms was proved, and numerical examples were provided to evaluate their performance. This final chapter summarizes the most significant results and outlines perspectives for future research.

In Chapter 3, we investigated the exact penalization of the incompressibility condition in the Bingham flow problem. In Theorem 3.1 and Remark 3.3 we established that the equivalence between the constrained and penalized formulations holds for all penalization parameters $\sigma > 0$, provided that $\sigma \geq \sigma_0$, where $\sigma_0 \approx \|\lambda\|_{L^2}|\Omega|^{\frac{1}{2}}$ and λ corresponds to the Lagrange multiplier. Although $\|\lambda\|_{L^2}$ cannot be computed *a priori* without an approximate solution, numerical experiments confirmed that the estimate of σ_0 is sharp.

The designed algorithm for the penalized Bingham problem utilizes second-order information to compute a descent direction without directly solving the minimum norm subgradient problem for the steepest descent direction. This simplification is justified by the analysis in Section 3.8.1, where it was discussed that the growth of the divergence of both the iterates and their associated descent directions is controlled by $\mathcal{O}(\gamma^{-1})$ on the active set A_γ^k at each iteration k. On the complement $\Omega \backslash A_\gamma^k$, we could only establish that this set vanishes numerically.

Numerical experiments demonstrated the benefits of the L^1-norm penalization of the divergence term. The exact penalization algorithm computes approximate solutions with a precise divergence. Notably, the L^1-norm promotes sparsification of the divergence term, as confirmed by the results in Section 3.9. Moreover, the experiments demonstrated the advantages of the L1-norm penalization for the divergence term compared to the SSN method. The SSN algorithm involves decoupling the underlying

system and stabilizing the equation for the velocity's divergence and the pressure p_r using a small parameter $\varsigma > 0$. This stabilization effectively relaxes the incompressibility constraint, which may lead to less accurate enforcement of the condition as shown in Table 3.2.

In Chapter 4, we developed an algorithm tailored for group-sparse problems, with applications to the simplified formulation of the Bingham flow and group-sparse PDE-constrained problems. For the first application, we designed an optimization algorithm to solve the inner problem of the augmented Lagrangian method (ALG1). This algorithm computes the steepest descent direction in a group-wise manner by evaluating the directional derivative, while second-order information is employed to precondition the search direction for improved efficiency and accuracy.

Additionally, the active-set strategy introduced in Section 4.2.5 for the Bingham flow problem is interpreted as a projection of the groups that are close to becoming sparse onto zero, enhancing sparsity detection. Compared to the *Alternating Direction Method of Multipliers* (ADMM or ALG2) in Table 4.1, the incorporation of the second-order information and the active-set phase improve the decrease of the cost functional and a slightly smaller constraint error. While the results of this first part are valuable from a theoretical perspective, our focus is on Section 4.3, where the methods devised for the group-sparse term are extended to address a more general optimization problem.

In this second section, we generalized the strategies devised for the Bingham flow in a pipe to design the Group-Sparse Descent Method (GSDM). We refine the strategies developed earlier by integrating the active-set strategy to reduce the dimensionality of the second-order system. Specifically, curvature information is applied exclusively to the non-sparse groups, thereby improving computational efficiency. The numerical experiment displayed in Table 4.5 shows that the number of unknown variables decreases throughout the execution of GSDM algorithm.

Furthermore, the projection phase is executed prior to constructing the descent direction, ensuring that sparsity is accounted for at an early stage of the algorithm. In Section 4.6.2 we discuss the effect of the active-set prediction phase in the semilineal ptimal control problem.

Through a local analysis of the index-sets involved in GSDM, we demonstrate in Theorem 4.9 that, near a local solution, the method simplifies to a Newton-type algorithm exhibiting quadratic convergence properties.

In addition, by focusing on the group structure of the problem, the method balances accuracy and computational cost, making it well-suited for a broader group-sparse optimization problems.

Finally, we outline potential directions for future research. The exact penalty algorithm developed for the incompressibility condition can be extended to other vis-

coplastic models, such as the Herschel-Bulkley and Casson models, which exhibit more complex rheological behaviors and are analyzed in Banach spaces. These generalizations could provide a broader framework for addressing a wider range of practical applications involving non-Newtonian fluids.

Another promising topic involves combining the exact penalty method with the GSDM to solve the Bingham flow problem, incorporating both nonsmooth terms: the L^1 -norm penalization of the incompressibility constraint and the unregularized term $g \int_\Omega |\mathcal{E}\mathbf{u}|\, dx$. This hybrid approach could offer a unified methodology for handling the coupled challenges of divergence-free constraints and yield-stress fluid behavior, providing a powerful tool for accurately modeling and simulating complex viscoplastic flows. Expanding these methods to more complex three-dimensional geometries or time-dependent problems also remains an open area of exploration.

Additionally, a further topic of interest could involve exploring adaptive strategies for estimating penalization parameters in group-sparse problems. This research could focus on developing methods to dynamically adjust the penalization parameters based on the problem's structure. Such strategies could provide more accurate solutions, particularly in high-dimensional or complex settings where the choice of parameters significantly impacts performance.

Bibliography

[1] Andrew & Gao, Scalable training of l1-regularized log-linear models, in: *Proceedings of the 24th international conference on Machine learning*, 2007, 33–40.

[2] Aposporidis, Haber, Olshanskii & Veneziani, A mixed formulation of the Bingham fluid flow problem: Analysis and numerical solution, *Computer methods in applied mechanics and engineering* 200 (2011), 2434–2446.

[3] Aravkin, Baraldi & Orban, A proximal quasi-Newton trust-region method for nonsmooth regularized optimization, *SIAM Journal on Optimization* 32 (2022), 900–929.

[4] Argyriou et al., Efficient first order methods for linear composite regularizers, *arXiv preprint arXiv:1104.1436* (2011).

[5] Arrow et al., *Studies in linear and non-linear programming*, Stanford University Press Stanford, 1958.

[6] Babuška, Error-bounds for finite element method, *Numerische Mathematik* 16 (1971), 322–333.

[7] Bagirov et al., *Numerical nonsmooth optimization*, Springer, 2020.

[8] Bagirov, Karmitsa & Mäkelä, *Introduction to Nonsmooth Optimization: theory, practice and software*, Springer, 2014.

[9] Bakin, Adaptive regression and model selection in data mining problems, PhD thesis, Mathematical Science Institute, The Australian National University, 1999.

[10] Baraldi & Kouri, A proximal trust-region method for nonsmooth optimization with inexact function and gradient evaluations, *Mathematical Programming* 201 (2023), 559–598.

[11] Bauschke, Combettes, et al., *Convex analysis and monotone operator theory in Hilbert spaces*, Springer, 2011.

[12] Bellavia, Malaspina & Morini, Inexact Newton methods with matrix approximation by sampling for nonlinear least-squares and systems, *arXiv preprint arXiv:2310.05501* (2023).

[13] Bender & Koschier, Divergence-Free SPH for Incompressible and Viscous Fluids, *IEEE Transactions on Visualization and Computer Graphics* 23 (2017), 1193–1206, DOI: 10.1109/TVCG.2016.2578335.

[14] Bercovier & Engelman, A finite-element method for incompressible non-Newtonian flows, *Journal of Computational Physics* 36 (1980), 313–326.

[15] Bertsekas & Mitter, Steepest descent for optimization problems with non-differentiable cost functionals, tech. rep., MIT, Dept. of Electrical Engineering, 1971.

[16] Bleyer, Maillard, Buhan & Coussot, Efficient numerical computations of yield stress fluid flows using second-order cone programming, *Computer Methods in Applied Mechanics and Engineering* 283 (2015), 599–614, DOI: 10.1016/J.CMA.2014.10.008.

[17] Bonnans, Gilbert, Lemaréchal & Sagastizábal, *Numerical optimization: theoretical and practical aspects*, Springer Science & Business Media, 2006.

[18] Bouchut, Eymard & Prignet, Convergence of conforming approximations for inviscid incompressible Bingham fluid flows and related problems, *Journal of Evolution Equations* 14 (2014), 635–669, DOI: 10.1007/S00028-014-0231-9.

[19] Boyd et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends® in Machine learning* 3 (2011), 1–122.

[20] Bresch, Fernandez-Nieto, Ionescu & Vigneaux, Augmented Lagrangian method and compressible visco-plastic flows: applications to shallow dense avalanches, in: *New directions in mathematical fluid mechanics*, Springer, 2009, 57–89.

[21] Byrd, Chin, Nocedal & Oztoprak, A family of second-order methods for convex ℓ_1—regularized optimization, *Mathematical Programming* (2012), 1–33.

[22] Casas, Herzog & Wachsmuth, Analysis of spatio-temporally sparse optimal control problems of semilinear parabolic equations, *ESAIM: Control, Optimisation and Calculus of Variations* 23 (2017), 263–295.

[23] Casas & Kunisch, Optimal control of semilinear elliptic equations in measure spaces, *SIAM Journal on Control and Optimization* 52 (2014), 339–364.

[24] CDC, Diabetes Health Indicators, Kaggle, 2015.

[25] Chambolle & Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of mathematical imaging and vision* 40 (2011), 120–145.

[26] Chauchat & Medale, A three-dimensional numerical model for incompressible two-phase flow of a granular bed submitted to a laminar shearing flow, *Computer*

Methods in Applied Mechanics and Engineering 199 (2010), 439–449, DOI: 10.1016/J.CMA.2009.07.007.

[27] Chen & Drapaca, H(div) conforming methods for the rotation form of the incompressible fluid equations, *Calcolo* 57 (2020), DOI: 10.1007/s10092-020-00380-8.

[28] Chen, Nashed & Qi, Smoothing methods and semismooth methods for nondifferentiable operator equations, *SIAM Journal on Numerical Analysis* 38 (2000), 1200–1216.

[29] Chouzenoux, Pesquet & Repetti, Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function, *Journal of Optimization Theory and Applications* 162 (2014), 107–132.

[30] Christof, De los Reyes & Meyer, A nonsmooth trust-region method for locally Lipschitz functions with application to optimization problems constrained by variational inequalities, *SIAM Journal on Optimization* 30 (2020), 2163–2196.

[31] Ciarlet, *Linear and nonlinear functional analysis with applications*, SIAM, 2013.

[32] Clarke, *Functional analysis, calculus of variations and optimal control*, Springer, 2013.

[33] Clarke, *Functional analysis, calculus of variations and optimal control*, Springer Science & Business Media, 2013.

[34] Clason & Valkonen, Introduction to nonsmooth analysis and optimization, *arXiv preprint arXiv:2001.00216* (2020).

[35] Cortez, Wine Quality, UCI Machine Learning Repository, 2009.

[36] De los Reyes & Merino, The second order method with enriched Hessian information for imaging composite sparse optimization problems, *preprint* (2020).

[37] De Los Reyes & González-Andrade, Path following methods for steady laminar Bingham flow in cylindrical pipes, *ESAIM: Mathematical Modelling and Numerical Analysis* 43 (2009), 81–117.

[38] De los Reyes & González-Andrade, Numerical simulation of two-dimensional Bingham fluid flow by semismooth Newton methods, *Journal of computational and applied mathematics* 235 (2010), 11–32.

[39] De Los Reyes, Loayza & Merino, Second-order orthant-based methods with enriched Hessian information for sparse ℓ_1-optimization, *Computational Optimization and Applications* 67 (2017), 225–258.

[40] De los Reyes, López-Ordóñez & Merino, A Second-order Method with Active-set Prediction for Group Sparce Optimization, *preprint* (2025).

[41] De Larrard & Roussel, Flow simulation of fresh concrete under a slipform machine, *Road Materials and Pavement Design* 12 (2011), 547–566.

[42] Dean, Glowinski & Guidoboni, On the numerical simulation of Bingham visco-plastic flow: old and new results, *Journal of non-newtonian fluid mechanics* 142 (2007), 36–62.

[43] Dennis Jr & Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, SIAM, 1996.

[44] Di Pillo & Grippo, An exact penalty function method with global convergence properties for nonlinear programming problems, *Mathematical Programming* 36 (1986), 1–18.

[45] DiBenedetto & Debenedetto, *Real analysis*, Springer, 2002.

[46] Duvant & Lions, *Inequalities in mechanics and physics*, Springer Science & Business Media, 2012.

[47] Ekeland & Temam, *Convex analysis and variational problems*, SIAM, 1999.

[48] Fortin & Glowinski, Méthodes de lagrangien augmenté: applications à la résolution numérique de problèmes aux limites, *(No Title)* (1982).

[49] Fortin & Glowinski, *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*, Elsevier, 2000.

[50] Frigaard & Ryan, Flow of a visco-plastic fluid in a channel of slowly varying width, *Journal of non-newtonian fluid mechanics* 123 (2004), 67–83.

[51] Fusi, Farina & Rosso, Retrieving the Bingham model from a bi-viscous model: some explanatory remarks, *Applied Mathematics Letters* 27 (2014), 11–14.

[52] Gabay & Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Computers & mathematics with applications* 2 (1976), 17–40.

[53] Geiger & Kanzow, *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer, 1999, URL: books.google.de/books?id=hzeklGTF4NEC.

[54] Giaquinta & Modica, *Mathematical analysis: An introduction to functions of several variables*, Springer Science & Business Media, 2010.

[55] Girault & Raviart, *Finite element methods for Navier-Stokes equations: theory and algorithms*, Springer Science & Business Media, 2012.

[56] Glowinski, Lions & Tremolieres, Numerical Analysis of Variational Inequalities, North-Holland, Amsterdam, 1981 ().

[57] Glowinski, *Numerical methods for nonlinear variational problems*, Springer Science & Business Media, 2013.

[58] Glowinski, On alternating direction methods of multipliers: a historical perspective, *Modeling, simulation and optimization for science and technology* (2014), 59–82.

[59] Glowinski & Le Tallec, *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*, SIAM, 1989.

[60] González-Andrade, Semismooth Newton and path following methods for the numerical simulation of Bingham fluids, PhD thesis, PhD thesis, Escuela Politecnica Nacional del Ecuador, 2008.

[61] González-Andrade & Lopez-Ordonez, A multigrid optimization algorithm for the numerical solution of quasilinear variational inequalities involving the p-Laplacian, *Computers & Mathematics with Applications* 75 (2018), 1107–1127.

[62] González-Andrade, López-Ordóñez & Merino, Nonsmooth exact penalization second-order methods for incompressible bi-viscous fluids, *Computational Optimization and Applications* 80 (2021), 979–1025.

[63] González-Andrade & Silva, H (div)-conforming and discontinuous Galerkin approach for Herschel–Bulkley flow with density-dependent viscosity and yield stress, *Applications in Engineering Science* 19 (2024), 100193.

[64] Herzog, Stadler & Wachsmuth, Directional sparsity in optimal control of partial differential equations, *SIAM Journal on Control and Optimization* 50 (2012), 943–963.

[65] Herzog, Stadler & Wachsmuth, Directional sparsity in optimal control of partial differential equations, *SIAM Journal on Control and Optimization* 50 (2012), 943–963.

[66] Hintermüller, Semismooth Newton methods and applications, *Department of Mathematics, Humboldt-University of Berlin* (2010).

[67] Hintermüller, Ito & Kunisch, The primal-dual active set strategy as a semismooth Newton method, *SIAM Journal on Optimization* 13 (2002), 865–888.

[68] Hinze, Pinnau, Ulbrich & Ulbrich, *Optimization with PDE constraints*, Springer Science & Business Media, 2008.

[69] Huber, Robust estimation of a location parameter, in: *Breakthroughs in statistics: Methodology and distribution*, Springer, 1992, 492–518.

[70] Huilgol, *Fluid mechanics of viscoplasticity*, Springer, 2015.

[71] Huilgol & You, Application of the augmented Lagrangian method to steady pipe flows of Bingham, Casson and Herschel–Bulkley fluids, *Journal of non-newtonian fluid mechanics* 128 (2005), 126–143.

[72] Huilgol & Nguyen, Variational principles and variational inequalities for the unsteady flows of a yield stress fluid, *International journal of non-linear mechanics* 36 (2001), 49–67.

[73] Ionescu, Augmented Lagrangian for shallow viscoplastic flow with topography, *Journal of Computational Physics* 242 (2013), 544–560.

[74] Ito & Kunisch, *Lagrange multiplier approach to variational problems and applications*, SIAM, 2008.

[75] Johannessen, Kumar & Kvamsdal, Divergence-conforming discretization for Stokes problem on locally refined meshes using LR B-splines, *Computer Methods in Applied Mechanics and Engineering* 293 (2015), 38–70, doi: 10.1016/J.CMA.2015.03.028.

[76] Kanzow & Lechner, Globalized inexact proximal Newton-type methods for nonconvex composite functions, *Computational Optimization and Applications* 78 (2021), 377–410.

[77] Kelley, *Iterative methods for linear and nonlinear equations*, SIAM, 1995.

[78] Kikuchi & Oden, *Contact problems in elasticity: a study of variational inequalities and finite element methods*, SIAM, 1988.

[79] Kummer, Generalized Newton and NCP-methods: Convergence, regularity, actions, *Discussiones Mathematicae, Differential Inclusions, Control and Optimization* 20 (2000), 209–244.

[80] Kummer et al., Newton's method for non-differentiable functions, *Advances in mathematical optimization* 45 (1988), 114–125.

[81] Laaber, *Numerical simulation of a three-dimensional Bingham fluid flow*, na, 2008.

[82] Lederer, Lehrenfeld & Schöberl, Hybrid Discontinuous Galerkin Methods with Relaxed H(div)-Conformity for Incompressible Flows. Part I, *SIAM J. Numer. Anal.* 56 (2017), 2070–2094, doi: 10.1137/17M1138078.

[83] Li & Lin, Accelerated proximal gradient methods for nonconvex programming, *Advances in neural information processing systems* 28 (2015).

[84] Li, Sun & Toh, A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems, *SIAM Journal on Optimization* 28 (2018), 433–458.

[85] Li et al., Manifold optimization-based analysis dictionary learning with an l12-norm regularizer, *Neural networks : the official journal of the International Neural Network Society* 98 (2018), 212–222, doi: 10.1016/j.neunet.2017.11.015.

[86] Lions, Optimal control of systems governed by partial differential equations, 1971.

[87] Lopes, Santos & Silva, Accelerating block coordinate descent methods with identification strategies, *Computational Optimization and Applications* 72 (2019), 609–640, doi: 10.1007/S10589-018-00056-8.

[88] Luenberger, Control problems with kinks, *IEEE Transactions on Automatic Control* 15 (1970), 570–575.

[89] Markelle Kelly & Nottingham, UCI Machine Learning Repository, UCI Machine Learning Repository, 2025, url: archive.ics.uci.edu.

[90] Mifflin, Semismooth and semiconvex functions in constrained optimization, *SIAM Journal on Control and Optimization* 15 (1977), 959–972.

[91] Moyers-Gonzalez & Frigaard, Numerical solution of duct flows of multiple visco-plastic fluids, *Journal of non-newtonian fluid mechanics* 122 (2004), 227–241.

[92] Muravleva, Uzawa-like methods for numerical modeling of unsteady viscoplastic Bingham medium flows, *Applied Numerical Mathematics* 93 (2015), 140–149, doi: 10.1016/J.APNUM.2014.06.001.

[93] Ng, Jordan & Weiss, On spectral clustering: Analysis and an algorithm, *Advances in neural information processing systems* 14 (2001).

[94] Noll, Convergence of non-smooth descent methods using the Kurdyka–Lojasiewicz inequality, *Journal of Optimization Theory and Applications* 160 (2014), 553–572.

[95] O'Donovan & Tanner, Numerical study of the Bingham squeeze film problem, *Journal of Non-Newtonian fluid mechanics* 15 (1984), 75–83.

[96] Ochs & Pock, Adaptive FISTA for nonconvex optimization, *SIAM Journal on Optimization* 29 (2019), 2482–2503.

[97] Papanastasiou, Flows of materials with yield, *Journal of rheology* 31 (1987), 385–404.

[98] Qi, Convergence analysis of some algorithms for solving nonsmooth equations, *Mathematics of operations research* 18 (1993), 227–244.

[99] Qi & Sun, A nonsmooth version of Newton's method, *Mathematical programming* 58 (1993), 353–367.

[100] Qin, Scheinberg & Goldfarb, Efficient block-coordinate descent algorithms for the group lasso, *Mathematical Programming Computation* 5 (2013), 143–169.

[101] Saramito & Roquet, An adaptive finite element method for viscoplastic fluid flows in pipes, *Computer methods in applied mechanics and engineering* 190 (2001), 5391–5412.

[102] Schirotzek, *Nonsmooth analysis*, Springer Science & Business Media, 2007.

[103] Schroeder & Lube, Divergence-Free H(div)-FEM for Time-Dependent Incompressible Flows with Applications to High Reynolds Number Vortex Dynamics, *Journal of Scientific Computing* 75 (2017), 830–858, doi: 10.1007/s10915-017-0561-1.

[104] Shapiro, On concepts of directional differentiability, *Journal of optimization theory and applications* 66 (1990), 477–487.

[105] Solntsev, Nocedal & Byrd, An algorithm for quadratic l1-regularized optimization with a flexible active-set strategy, *Optimization Methods and Software* 30 (2015), 1213–1237.

[106] Stadler, Elliptic optimal control problems with L 1-control cost and applications for the placement of control devices, *Computational Optimization and Applications* 44 (2009), 159–181.

[107] Sverdrup, Nikiforakis & Almgren, Highly parallelisable simulations of time-dependent viscoplastic fluid flow with structured adaptive mesh refinement, *Physics of Fluids* (2018), doi: 10.1063/1.5049202.

[108] Tanner & Milthorpe, Numerical simulation of the flow of fluids with yield stress, *Numer Meth Lami Turb Flow Seattle* (1983), 680–690.

[109] Temam, *Navier–Stokes equations: theory and numerical analysis*, American Mathematical Society, 2001.

[110] Trémolieres, Lions & Glowinski, *Numerical analysis of variational inequalities*, Elsevier, 2011.

[111] Treskatis, Fast proximal algorithms for applications in viscoplasticity. (2016).

[112] Treskatis, Moyers-González & Price, An accelerated dual proximal gradient method for applications in viscoplasticity, *Journal of Non-Newtonian fluid mechanics* 238 (2016), 115–130.

[113] Tröltzsch, *Optimal control of partial differential equations*, American Mathematical Society, 2010, xvi+399.

[114] Ulbrich, Semismooth Newton methods for operator equations in function spaces, *SIAM Journal on Optimization* 13 (2002), 805–841.

[115] Ulbrich, *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, SIAM, 2011.

[116] Vaidya et al., Channel flow of MHD bingham fluid due to peristalsis with multiple chemical reactions: an application to blood flow through narrow arteries, *SN Applied Sciences* 3 (2021), 1–12.

[117] Vola, Boscardin & Latché, Laminar unsteady flows of Bingham fluids: a numerical strategy and some benchmark results, *Journal of Computational Physics* 187 (2003), 441–456.

[118] Von Luxburg, A tutorial on spectral clustering, *Statistics and computing* 17 (2007), 395–416.

[119] Wachs, Numerical simulation of steady Bingham flow through an eccentric annular cross-section by distributed Lagrange multiplier/fictitious domain and augmented Lagrangian methods, *Journal of Non-newtonian Fluid Mechanics* 142 (2007), 183–198, doi: 10.1016/J.JNNFM.2006.08.009.

[120] Wilbrandt, *Stokes–Darcy Equations: Analytic and Numerical Analysis*, Springer, 2019.

[121] Yousefpour, Combination of steepest descent and BFGS methods for nonconvex nonsmooth optimization, *Numerical Algorithms* 72 (2016), 57–90.

[122] Yu, Vishwanathan, Günter & Schraudolph, A quasi-Newton approach to non-smooth convex optimization, in: *Proceedings of the 25th international conference on Machine learning*, 2008, 1216–1223.

[123] Yuan & Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68 (2006), 49–67.

[124] Zhang, An augmented Lagrangian approach to Bingham fluid flows in a lid-driven square cavity with piecewise linear equal-order finite elements, *Computer Methods in Applied Mechanics and Engineering* 199 (2010), 3051–3057, doi: 10.1016/J.CMA.2010.06.020.

[125] Zhang, Physics-Informed Neural Networks for Bingham Fluid Flow Simulation Coupled with an Augmented Lagrange Method, *AppliedMath* (2023), doi: 10.3390/appliedmath3030028.

[126] Zowe & Kurcyusz, Regularity and stability for the mathematical programming problem in Banach spaces, *Applied mathematics and Optimization* 5 (1979), 49–62.