UNIVERSITÄT
BAYREUTH

INEXACT PROXIMAL NEWTON METHODS FOR FINITE STRAIN PLASTICITY

Von der Universität Bayreuth
zur Erlangung des Grades eines

Doktors der Naturwissenschaften
(Dr. rer. nat.)

genehmigte Abhandlung

von

Bastian Hermann Pötzl

geboren in Tirschenreuth

1. Gutachter:	Prof. Dr. Anton Schiela
2. Gutachter:	Prof. Dr. Christian Clason
Tag der Einreichung:	06.03.2023
Tag des Kolloquiums:	19.07.2023

Zusammenfassung

Energetische Formulierungen der Plastizität endlicher Verzerrungen sind ein Beispiel für die allgemeine Theorie der ratenunabhängigen Systeme. Sie sind als direkte Verallgemeinerung der primalen Formulierung der Plastizität kleiner Verzerrungen zu sehen, bei der die Unbekannten als die Verschiebungen, die plastische Dehnung und möglicherweise Verhärtungsvariablen gegeben sind. Insbesondere benutzt der Begriff energetischer Lösungen keine Ableitungen, was die elegante Modellierung nicht-glatter Phänomene ermöglicht.

Weiterhin können energetische Lösungen als Grenzwerte stückweise konstanter Interpolanten von Lösungen für zeitinkrementelle Minimierungsprobleme charakterisiert werden, wodurch die zugehörige Formulierung zugänglich für Optimierungsalgorithmen wird. Allerdings beinhalten die zeitinkrmentellen Minimierungsprobleme einige Schwierigkeiten: Sie sind hochgradig nicht-linear, nicht-konvex und nicht-glatt. Die Entwicklung entsprechender Lösungsalgorithmen ist das Ziel der vorliegenden Arbeit.

Nach dem Erklären der speziellen Problemstruktur sowohl im allgemeinen Rahmen ratenunabhängier Systeme als auch als konkrete Formulierung für die Plastizität endlicher Verzerrungen wird die algorithmische Idee von Proximal Newton Methoden aus der Literatur für endlichdimensionale Optimierung vorgestellt. Da bestehende Formulierungen nicht die Behandlung von Funktionenraumproblemen erlauben, werden algorithmische Konzepte und Konvergenztheorie auf ein hinreichend allgemeines Hilbertraumszenario angepasst.

Während die Differenzierbarkeitsbedingungen mithilfe sowohl bekannter als auch neuartiger Semiglattheitsbegriffe gelockert werden, hilft eine quadratische Normregularisierung im Subproblem zur Schrittberechnung dabei, restriktive Konvexitätsannahmen an das Zielfunktional zu beseitigen. Globale Konvergenz und lokale Beschleunigung der Proximal Newton Methode werden bewiesen und numerische Robustheit nahe an optimalen Lösungen wird durch das Einführen eines alternativen Abstiegskriteriums für Szenarios, die anfällig für numerische Auslöschung sind, gesichert.

Die inexakte Berechnung von Schritten und adaptive Strategien zur Parameterwahl verbessern die Effizienz der Berechnungen im Algorithmus noch weiter – und zwar unter Aufrechterhaltung der vorteilhaften Konvergenzeigenschaften. Insbesondere müssen dabei auch die zugehörigen Inexaktheitskriterien für die effiziente Auswertung im Funktionenraum konzipiert werden. Der Einfluss dieser algorithmischen Modifikationen wird numerisch anhand einer Reihe von Modellproblemen im Funktionenraum untersucht.

Zuletzt wird die finale Form des Lösungsalgorithmus auf ein anspruchsvolles und realistisches Anwendungsproblem aus dem Bereich der Plastizität endlicher Verzerrungen angewendet: die Deformation einer stahl-ähnlichen Büroklammer in einem binären Homotopieproblem, das aus dem Laden mit verschieden starken Kräften und dem nachfolgenden Entladen besteht, um die irreversible Natur plastischer Verformungen zu demonstrieren.

Abstract

The energetic formulation of finite strain plasticity is an instance of the general theory of rate-independent systems. It can be understood as a direct generalization of the primal formulation of small strain plasticity where the unknowns are the displacements, plastic strains, and possibly hardening variables. In particular, the notion of energetic solutions does not involve derivatives which allows for modeling non-smooth phenomena in an elegant way.

Furthermore, energetic solutions can be characterized as the limits of piecewise constant interpolants of solutions to time-incremental minimization problems which makes the corresponding formulation amenable to optimization algorithms. However, various difficulties are present in the time-incremental minimization problems: They are highly non-linear, non-convex, and non-smooth. The development of adequate solution algorithms is the goal of the present treatise.

After the particular problem structure is presented both in the general framework of rate-independent systems and in the concrete formulation of finite strain plasticity, the algorithmic idea of Proximal Newton methods for composite minimization problems is introduced as considered in the literature for finite dimensional optimization. Since existing formulations do not allow for the treatment of function space problems like finite strain plasticity, algorithmic concepts and convergence theory are adapted to a sufficiently general Hilbert space scenario.

While the framework of differentiability is loosened by developing a formulation which adequately uses both known and novel concepts of semi-smoothness, restrictive convexity assumptions on the composite objective functional are eliminated by quadratic norm regularization in the update step computation subproblems. Global convergence and local acceleration of the Proximal Newton method are established and numerical robustness close to optimal solutions is ensured by introducing an alternative sufficient decrease criterion for scenarios susceptible to numerical cancellation.

Inexact computation of update steps and adaptive strategies for choosing algorithmic parameters further improve the computational efficiency of the algorithm while preserving advantageous convergence properties. In particular, also the corresponding inexactness criteria for update step candidates are designed for efficient evaluation in a function space scenario. The influence of these algorithmic modifications is investigated numerically in a series of function space model problems.

Lastly, the final form of the solution algorithm is exposed to a demanding real world application problem from the framework of finite strain plasticity: the deformation of a steel-like paperclip in a binary homotopy problem which consists of loading with forces of different intensity and unloading in order to showcase the irreversible nature of plastic deformations.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Anton Schiela, for giving me the chance to pursue mathematical research in this both challenging and exciting field. Being new to the topic of non-linear optimization, his constructive feedback and open ear for questions of any kind were instrumental in shaping my work, and I could not have completed this project without his help. Even though our ways might part here for now, I will always remember him not only as an incredibly talented – though sometimes "dispersed" – mathematician but also as an empathetic and amicable person whom I am glad to have gotten to know.

Thanks go to Patrick Jaap for accompanying me throughout the journey of the last three and a half years. I could not have imagined a better project partnership, both from a social and a professional perspective.

I would also like to thank Christian Clason for committing his time to refereeing this thesis.

Lastly, I am extremely grateful for the support of my family, friends, and girlfriend. Their continuous distraction from my research made the past few years the best of my life.

This work was funded by the DFG SPP 1962: Non-smooth and Complementarity-Based Distributed Parameter Systems – Simulation and Hierarchical Optimization; Project Number: SCHI 1379/6-1.

"*Drauf angesetzt, Perfektionist.*
Bestand den Test, der letzte Schliff.
Der ganze Stress, er rechnet sich.
Mein Manifest, mein bestes Ich."

– Umse, Ohne Titel (2022)

Contents

Chapter 1

Introduction

1.1 A Historical Overview of Approaches Towards Mathematical Elasto-Plasticity

The mathematical modeling of elasto-plastic problems is essential for a wide range of fields, including structural design, manufacturing processes, technological production, and – of course – scientific research. Under specific conditions, several material types such as metals, concrete, rocks, clays, and soils may in fact be considered as plastic. While the origin of plasticity can be traced back to the mid-nineteenth century to the work of Tresca [105], more contributions to the classical theory of plasticity started to appear around the first half of the twentieth century, cf. [39, 46, 71, 82, 98].

Within the last decades, a considerable amount of progress has been made in the theory of elasto-statics at finite strains. Phenomena inherent to the latter field exhibit geometric non-linearities as well as physically necessary singularities which is why the approaches for their treatment developed in [5] have been considered a breakthrough in the field. Approximately at the same time, also the theory of elasto-plasticity enjoyed major attention which have led to a rigorous mathematical basis, cf. [1, 36, 73, 103]. In these early developments, however, authors mainly restricted their deliberations to the case of small strains where methods of convex analysis in Hilbert spaces can be taken advantage of straight-forwardly.

For the consideration of elasto-plasticity at finite strains, the authors in [78, 79] established the fact that time-incremental problems in the rate-independent and in the viscoplastic case can be written as a minimization problem for the sum of the increments in the stored energy and the dissipated energy. In particular, this major advance allowed for the proof of general existence results for the time-continuous problem by direct methods in the calculus of variations, cf. [60]. Furthermore, the theory developed within the latter work takes advantage of the concept of energetic solutions for rate-independent systems which have first been considered in [67] and further developed in [30, 62, 66]. This notion allows for coping with the strong non-linearities generated by the particular Lie group structure within the general and special linear matrix group.

Even though also efficient numerical methods have been developed and successfully implemented, cf. [93], neither of these schemes has been supported by rigorous convergence analysis up until [65]. There, recent advances in abstract numerical approaches for rate-independent processes [64, 66] are used in order to construct specific finite-element numerical schemes for gradient plasticity at large strains and guarantee their convergence in sufficient generality. Even there, however, the question of how to solve the occurring time-incremental minimiza-

tion problems efficiently remains somewhat vague. We will cover this peculiar topic within the present treatise by investigating a whole class of minimization algorithms which can in particular be utilized for that task.

1.2 Contributions and Outline

As mentioned above, the main objective of this thesis is the development of an efficient function space minimization algorithm which can cope with the structural peculiarities exhibited by time-incremental minimization problems stemming from finite strain plasticity formulated in the framework of rate-independent systems. A thorough understanding of both the problem structure in function space and existing approaches towards corresponding minimization algorithms is established in order to then design an iterative method which features both algorithmic functionality and computational effectiveness. The structure is as follows:

Chapter 1 – Introduction. Now that we have given a short historical overview of mathematical approaches for elasto-plasticity, the remainder of the chapter provides a collection of standard notation which is used throughout the present treatise. Most of the particular formulations, however, will be explained when they first appear in the main part of the manuscript.

Chapter 2 – Rate-Independent Finite Strain Plasticity. This chapter introduces the motivational application problem for the development of our function space algorithm: time-incremental minimization problems in the framework of finite strain plasticity. In order to adequately formulate the particular problem setting and structure, we first elaborate on the general concept of rate-independent systems together with the corresponding problem formulations, solution concepts, and existence theory of solutions. Afterwards, we review the physical description of finite strain plasticity and fit it into the framework of rate-independent systems. Lastly, we formulate the specific form of the application problem as we will solve it later on in our numerical investigations.

Chapter 3 – Proximal Newton Methods in Hilbert Spaces. With the specific problem formulation at hand, we depart on the endeavor of developing an adequate solution algorithm. To this end, we first review existing approaches for minimizing objective functionals – mainly in finite dimensional scenarios. In that process, we emphasize central principles of non-linear optimization which we will come back to in the design of our algorithm. Afterwards, we lay the foundation for the final formulation of the algorithm by generalizing existing ideas for Proximal Newton methods to an infinite dimensional Hilbert space setting. The particular structure of our method allows for the application to problems with objective functions exhibiting rather inconvenient differentiability and convexity properties.

Chapter 4 – Modifications for Algorithmic Efficiency. Even though the previously constructed function space algorithm constitutes a functioning method applicable to problems of the desired nature, the aspect of computational efficiency has been disregarded up to that point. For this reason, we develop inexactness criteria for update step candidates of the Proximal Newton algorithm which allow to both save computational time and preserve convergence properties of the exact method. Particular importance is laid on the efficient evaluation of

these inexactness criteria in the infinite dimensional Hilbert space setting in which we want to apply our method. Furthermore, we explore possibilities to adaptively choose algorithmic parameters over the course of our algorithm in order to take advantage of the structure of the underlying minimization problem even further.

Chapter 5 – Application to Finite Strain Plasticity. The modified version of our solution algorithm is compared with its unmodified variant in the computation of solutions for a series of demanding time-incremental minimization problems stemming from the formulation of finite strain plasticity in the framework of rate-independent systems. This allows for an assessment to which extent the modifications developed in the previous chapter improve the algorithmic efficiency of our method. Afterwards, the modified version of the algorithm is employed in order to solve yet another finite strain plasticity problem with a more complex problem geometry and yet another material model governing the response of the test object. The latter is given by a metal paperclip and thus represents an illustrative example for finite strain plasticity theory in everyday life.

Chapter 6 – Conclusion. The results of this thesis are summarized and put in perspective with respect to open questions and current research on related topics.

Appendix. The appendix contains specifications on the implementation which are not elaborately explained in the main part of the manuscript. Additionally, we list the test machine specifications and describe the availability of computational data from our numerical investigations.

1.3 Some Notation

Let us here give a short overview over some standard notation which we will use across most chapters in the following. While chapter- or even section-specific notation will be introduced just where it is needed, we will cover some more general concepts here. Furthermore, a list of symbols is included after the appendix.

Spaces and Norms

To make a start, the natural and real numbers are denoted by the symbols $\mathbb{N}$ and $\mathbb{R}$. Let us then consider some $d \in \mathbb{N}$ together with the corresponding d-dimensional Euclidean space $\mathbb{R}^d$. By

$$\left\|\cdot\right\| : \mathbb{R}^d \to \mathbb{R} \quad \text{and} \quad \langle \cdot, \cdot \rangle_{\mathbb{R}^d} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

we refer to any of the equivalent norms and the Euclidean scalar product. Obviously, these concepts can be transferred to more general spaces X where we will then write $\left\|\cdot\right\|_X$ for a fixed one of the generally non-equivalent norms, and $\langle \cdot, \cdot \rangle_X$ in case a scalar product is defined, accordingly. On Hilbert spaces $(X, \langle \cdot, \cdot \rangle_X)$, the norm is always induced by the scalar product $\langle \cdot, \cdot \rangle_X$ unless otherwise stated.

As far as notable subsets of the general matrix space $\mathbb{R}^{d\times d}$ are concerned, we will often use the *general linear group* $\mathrm{GL}^+(d)$ and, as a subset of the former, also the *special linear group*

$\mathrm{SL}(d)$ defined by

$$\mathrm{GL}^+(d) \coloneqq \left\{ \mathrm{M} \in \mathbb{R}^{d\times d} \,\middle|\, \det(\mathrm{M}) > 0 \right\} \quad \text{and} \quad \mathrm{SL}(d) \coloneqq \left\{ \mathrm{M} \in \mathbb{R}^{d\times d} \,\middle|\, \det(\mathrm{M}) = 1 \right\}. \tag{1.3.1}$$

Additionally, we write $\mathrm{I} \in \mathbb{R}^{d\times d}$ for the identity matrix.

Metric spaces (X, d_X) endowed with some metric $d_X \colon X \times X \to \mathbb{R}$ allow us to define the ball of radius r around an element $x \in X$ by

$$B_r(x) \coloneqq \left\{ y \in X \,\middle|\, d_X(x,y) \leq r \right\}.$$

The space of continuous linear mappings between normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ is denoted by $\mathcal{L}(X,Y)$ which also lets us introduce the dual space of X as $X^* \coloneqq \mathcal{L}(X, \mathbb{R})$. Then, the evaluation of a linear functional $\varphi \in X^*$ for some $v \in X$ is either emphasized by writing $\langle \varphi, v \rangle$ or simply denoted by φv in qualified situations for the sake of notational simplicity.

Differentiation, Integration, and Function Spaces

Lastly, let us shortly elaborate on the notation regarding differentiation, integration, and the ensuing function spaces. By D_v we refer to the partial (weak) derivative of some functional with respect to the placeholder variable v within its definition. The dot $\dot{\mathbf{P}}$ signifies – depending on the context of its use – either the (weak) time-derivative of the corresponding time-dependent functional or the (weak) derivative of a parameterized curve with respect to the corresponding parameter within its definition.

Let us now consider some function $\phi \colon X \times V \to \mathbb{R}$, $(x,v) \mapsto \phi(x,v)$ on general Banach spaces X, V. In case ϕ is convex concerning the placeholder variable v, the *convex subdifferential* with respect to v at any $(x,v) \in X \times V$ is referred to as

$$\partial^v \phi(x,v) \coloneqq \left\{ \xi \in V^* \,\middle|\, \forall \tilde{v} \in V \colon \phi(x,\tilde{v}) \geq \phi(x,v) + \langle \xi, \tilde{v} - v \rangle \right\} \tag{1.3.2}$$

where the variable indicator is omitted in case ϕ only depends on one variable. The above set of subderivatives can be generalized to mappings $\psi \colon X \times V \to \mathbb{R}$, $(x,v) \mapsto \psi(x,v)$, which are non-convex also in the v-variable with the use of the *Fréchet-subdifferential*

$$\partial_F^v \psi \coloneqq \left\{ \xi \in V^* \;\middle|\; \liminf_{u \to v} \frac{\psi(x,u) - \psi(x,v) - \langle \xi, u - v \rangle}{\|u - v\|_V} \geq 0 \right\} \tag{1.3.3}$$

which coincides with $\partial^v \psi$ in case ψ is convex with respect to v, cf. [51]. Also here, we omit the variable indicator in the trivial case. The definitions of the generalized differentials ∂_B and ∂_G will be given and elaborated on in Section 3.1.2.

As far as integration is concerned, both our notation and setup are straight-forward and easy to understand. All appearing integrals are to be understood with respect to the corresponding d-dimensional Lebesgue measure $\mathscr{L}^d$ where d depends on the variable to be integrated over. With this specification out of the way, for any subset $\Omega \subset \mathbb{R}^d$ we can introduce the *Lebesgue spaces* $L^p(\Omega)$ for $1 \leq p \leq \infty$. These are defined as equivalence classes of (extended) real valued functions defined on Ω that are finitely p-integrable, or essentially bounded with respect to the Lebesgue measure when $p = \infty$. The standard norms are

$$\|v\|_{L^p(\Omega)} \coloneqq \Big(\int_\Omega |v(x)|^p \mathrm{d}x \Big)^{\frac{1}{p}} \quad \text{and} \quad \|v\|_{L^\infty(\Omega)} \coloneqq \operatorname*{ess\,sup}_{x \in \Omega} |v(x)|$$

for $p < \infty$ and $p = \infty$ respectively. When elements of these spaces are referred to as functions, this is understood to mean the entire class of functions.

For an open set $\Omega \in \mathbb{R}^d$, the *Sobolev space* $W^{k,p}(\Omega)$ contains all functions in $L^p(\Omega)$ with finitely p-integrable weak derivatives up to order $k \in \mathbb{N}$, i.e.,

$$W^{k,p}(\Omega) \coloneqq \left\{ v \in L^p(\Omega) \,\middle|\, \forall |\alpha| \leq k \colon \mathrm{D}^\alpha v \in L^p(\Omega) \right\}$$

where $|\alpha| \coloneqq \sum_{k=0}^{d} \alpha_k$ is the order of a multi-index $\alpha = (\alpha_1, ..., \alpha_d) \in \left(\mathbb{N} \cup \{0\}\right)^d$ and $\mathrm{D}^\alpha v$ denotes the corresponding mixed partial α-derivative. The standard norms of some $v \in W^{k,p}(\Omega)$ are given by

$$\|v\|_{W^{k,p}(\Omega)} \coloneqq \Big(\sum_{0 \leq |\alpha| \leq k} \|\mathrm{D}^\alpha v\|_{L^p(\Omega)}^p \Big)^{\frac{1}{p}} \quad \text{and} \quad \|v\|_{W^{k,\infty}(\Omega)} \coloneqq \max_{|\alpha| \leq k} \|\mathrm{D}^\alpha v\|_{L^\infty(\Omega)} .$$

For bounded strong Lipschitz domains $\Omega \subset \mathbb{R}^d$, the (surjective) *trace operator* for the boundary $\partial\Omega$ is

$$\mathrm{tr} \colon W^{1,p}(\Omega) \to W^{1-\frac{1}{p},p}(\partial\Omega) .$$

With this definition at hand, we can then consider

$$W_0^{1,p}(\Omega) \coloneqq \left\{ v \in W^{1,p}(\Omega) \,\middle|\, \mathrm{tr}(v) = 0 \text{ a.e. on } \partial\Omega \right\}$$

where "a.e. on $\partial\Omega$" is to be understood in sense of the $(d-1)$-dimensional Hausdorff measure on $\partial\Omega$. For $p = 2$ and $k \in \mathbb{N}$, the sets as defined above are Hilbert spaces and the standard abbreviation $H^k(\Omega) \coloneqq W^{k,2}(\Omega)$ together with $H_0^k(\Omega) \coloneqq W_0^{k,2}(\Omega)$ is applied.

Chapter 2

Rate-Independent Finite Strain Plasticity

The aim of this chapter consists of providing sufficient background knowledge for the main application of our work, namely describing plastic deformations at finite strains. To this end, in Section 2.1 we motivate, introduce, and study so-called *rate-independent systems* as considered in [66]. These will constitute a fitting framework for investigating the finite strain plasticity problem and provide an adequate concept of solutions together with an intuitive way to approximately compute them numerically by solving a series of so-called *homotopy step problems*. Later on, in Section 2.2, we deduce the elasto-plasticity problem at finite strains from continuum mechanics and present a formulation of it fitting into the framework of rate-independent systems. At last, we introduce and simplify the particular application problem which we will later on consider in Chapter 5 for our numerical simulations.

2.1 Rate-Independent Systems

As mentioned above, we will now bring the mathematical theory for continuum mechanics of elastic solids into a more general framework which is also capable of coping with for example poling-induced piezoelectricity or viscodynamics: so-called *rate-independent systems*. The theory behind these has been thoroughly investigated in [66] and the following section of the present treatise is mainly a concise reformulation of selected contents presented there with some additional elaborations and filled-in gaps.

We will motivate and define the notion of rate-independent systems in Section 2.1.1 and afterwards give different approaches to formulating the general problem as well as provide concepts of solutions to the latter in Section 2.1.2. Finally, in Section 2.1.3, we will give an intuitively accessible version of the requirements for and motivate their contribution to the proof of the central existence result for solutions of rate-independent systems – in particular with regard to its application in the later stages of the present treatise.

2.1.1 Motivation and Definition

A central notion for the motivation and understanding of the concept of rate-independent systems is the one of intrinsic time scales and their relation to each other. In order to on the one hand understand the nature of intrinsic time scales and give a first and rather accessible example of a rate-independent system, we will consider the following type of ordinary

differential equations arising in mechanics, cf. [66, Chapter 1]:

$$\mathrm{M}\ddot{q} + F(\dot{q}) + \mathrm{K}q = \hat{\ell}(t)\,. \tag{2.1.1}$$

The differential equation describes the evolution of the state $q : [0,T] \to \mathbb{R}^n$ for some $T > 0$ and $n \in \mathbb{N}$. The symmetric and positive definite matrices $\mathrm{M}, \mathrm{K} \in \mathbb{R}^{n\times n}$ represent the respective mass and stiffness tensor of the underlying system. The vector-field $F : \mathbb{R}^n \to \mathbb{R}^n$ describes the influence of (possibly non-linear) damping while $\hat{\ell} : [0,T] \to \mathbb{R}^n$ is a time-dependent external force.

Assuming for now that the damping is governed by some symmetric viscosity matrix V via $F(\dot{q}) = \mathrm{V}\dot{q}$, we can characterize the aforementioned time-scales of the described system as follows: Firstly, internal time scales which split up to the *dynamic time scale* corresponding to oscillatory frequencies (eigenvalues of $\mathrm{M}^{-1}\mathrm{K}$) and the *viscous time scale* corresponding to viscous relaxation rates (eigenvalues of $\mathrm{V}^{-1}\mathrm{K}$) in contrast to, secondly, the *time scale of external loading* which can be characterized by $\left\|\frac{\mathrm{d}}{\mathrm{d}t}\hat{\ell}(t)\right\|^{-1}$.

The concept of *rate-independence* now delineates the case where internal processes develop significantly faster than the speed at which external loading takes place. This idealization happens in the limit case for the respective time scale of external loading. In order to illustrate this behavior in the framework of the above mechanical ordinary differential equation, we consider a slowly varying load $\hat{\ell}_\varepsilon(t) = \ell(\varepsilon t)$ for some small $\varepsilon > 0$ and fixed $\ell : [0,T] \to \mathbb{R}^n$. This reparameterization will help us manipulate the respective time scales within the system.

For better characterization of the ensuing equation, we choose a specific power-law friction of the form $F(\dot{q}) = \nu|\dot{q}|^{\alpha-1}\dot{q}$ for the non-linear damping term with $\alpha, \nu > 0$. Since it is our goal to keep the external loading time scale constant within the limit process, we rescale the time variable by $\varepsilon t \to t$ and obtain

$$\varepsilon^2\mathrm{M}\ddot{q} + \varepsilon^\alpha\nu|\dot{q}|^{\alpha-1}\dot{q} + \mathrm{K}q = \ell(t)$$

where the interesting scenario of $\alpha \to 0$ and (afterwards) $\varepsilon \to 0$ provides us with the differential inclusion

$$\partial\mathcal{R}(\dot{q}) + \mathrm{K}q \ni \ell(t) \tag{2.1.2}$$

for the so-called *dissipation potential* $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}$, $\mathcal{R}(w) \coloneqq \nu|w|$.

A consequence of the limit process of $\varepsilon \to 0$ after rescaling $\varepsilon t \to t$ is the deterioration of the internal time scales of the system. Thus, this limit system suffices our first idea for the notion of rate-independence. Another helpful property which we will keep in mind for later generalizations is the following: Every pair of some state q and external load ℓ solving (2.1.2) can be rescaled to another pair $(\tilde{q}, \tilde{\ell})$ with $\tilde{q}(t) = q(\lambda t)$ and $\tilde{\ell}(t) = \ell(\lambda t)$ which then again also solves (2.1.2) for any $\lambda > 0$.

General Rate-Independence and Characterization by Homogeneity

Our original, rather ideological characterization of rate-independence still has one quite unpractical peculiarity hindering us to use it as a rigorous definition for the concept: Due to the dependence on the relation of internal and external time scales it can only cope with systems driven by a time-dependent loading input, thus restricting us to the area of non-autonomous evolutionary systems. Additionally, internal time scales are rarely easy to determine.

For this reason, we will now allow for a more general time dependence within the differential inclusion characterizing our system than just via $\ell\colon [0,T] \to \mathbb{R}^n$. In order to illustrate this

ambition of ours, we again consider (2.1.2) on some more general *state space* Q and define the *stored energy functional*

$$\mathcal{E}\colon [0,T]\times Q\to\mathbb{R}\,,\quad \mathcal{E}(t,q)\coloneqq\frac{1}{2}\langle q,\mathrm{K}q\rangle_Q-\langle \ell(t),q\rangle_Q\,. \tag{2.1.3}$$

Furthermore, we also generalize the dissipation potential $\mathcal{R}\colon Q\to\mathbb{R}$ to now be possibly state dependent via $\mathscr{R}: Q\times Q\to[0,\infty]$, $(q,v)\mapsto\mathscr{R}(q,v)$. More specifically, we assume that the general dissipation potential suffices

$$\mathscr{R}(q,\cdot)\colon Q\to[0,\infty] \text{ is convex and lower semi-continuous}\,,\ \mathscr{R}(q,0)=0 \tag{2.1.4}$$

for all states $q\in Q$. With these prerequisites at hand, the differential inclusion (2.1.2) takes the form

$$0\in\partial^v\mathscr{R}\big(q(t),\dot q(t)\big)+\mathrm{D}_q\mathcal{E}\big(t,q(t)\big) \tag{2.1.5}$$

for any $t\in[0,T]$. We can give an illustrative interpretation of (2.1.5) in the form of a force balance in Q^* where the dissipative force $\partial^v\mathscr{R}\big(q(t),\dot q(t)\big)$ must equilibrate the potential restoring force $-\mathrm{D}_q\mathcal{E}\big(t,q(t)\big)$ at time $t\in[0,T]$.

The formulation via (2.1.5) allows us to now define a suitable *solution mapping* if we additionally prescribe some time interval together with a suitable initial condition. This then leads to a complete description of the underlying system by referring to the following definition:

Definition 2.1.1: Rate-Independent Systems

Consider a time interval $[0,T]$ together with some Banach Space Q for the state, an energy functional $\mathcal{E}\colon[0,T]\times Q\to\mathbb{R}$ and a dissipation potential $\mathscr{R}\colon Q\times Q\to[0,\infty]$ satisfying (2.1.4). For $[t_1,t_2]\in\mathcal{J}_{[0,T]}$ and $q_1\in Q$, we define the **solution mapping**

$$\begin{aligned}&\mathcal{S}_{(Q,\mathcal{E},\mathcal{R})}\colon \mathcal{J}_{[0,T]}\times Q\to \mathrm{W}^{1,1}\big(0,T;Q\big)\,,\\ &\big([t_1,t_2],q_1\big)\mapsto\big\{q\colon[t_1,t_2]\to Q\;\big|\;q(t_1)=q_1\,,\,\forall_{a.a.}t\in[t_1,t_2]\colon (2.1.5)\text{ holds}\big\}\,.\end{aligned}$$

Then, the triple $(Q,\mathcal{E},\mathcal{R})$ is called a **rate-independent system** if for every $\big([t_1,t_2],q_1\big)\in J_{[0,T]}\times Q$ and every strictly increasing continuous time reparameterization $\alpha\colon[t_1,t_2]\to[t_1^*,t_2^*]\in\mathcal{J}_{[0,T]}$ with $\alpha(t_i)=t_i^*$, $i=1,2$, the following equivalence holds:

$$q\in\mathcal{S}_{(Q,\mathcal{E},\mathcal{R})}\big([t_1,t_2],q_1\big)\quad\Leftrightarrow\quad q\circ\alpha\in\mathcal{S}_{(Q,\mathcal{E},\mathcal{R})}\big([t_1^*,t_2^*],q_1\big)$$

Remark. *Note that the above definition of solutions for rate-independent systems considers the Sobolev space $W^{1,1}\big(0,T;Q\big)$ which is, also in the Banach space valued case, continuously embedded into $C\big([0,T];Q\big)$, cf. e.g. [27, Section 5.9.2, Theorem 2]. Thus, imposing the initial condition $q(t_1)=q_1\in Q$ is legitimate. The respective state equation* (2.1.5)*, however, can only be satisfied almost everywhere on $[t_1,t_2]$.*

Let us point out here that in [66] rate-independent systems also have to satisfy two additional properties, namely the so-called *concatenation* and *restriction* property. The concatenation property states that merging two established solutions in time yields a solution for the longer in-time problem. The restriction property states the opposite situation where restricted

solutions are still solutions to the problem on a subset of $[0,T]$. Since these properties do not play a role for the existence of solutions and determining them numerically in our context, we drop them here for the sake of simplicity.

Before we now give an insight to more accessible problem formulations and solution concepts for rate-independent systems, we will formulate a sufficient condition for the dissipation potential $\mathscr{R}$ such that the ensuing $(Q,\mathcal{E},\mathscr{R})$ is rate-independent.[1] To this end, we introduce the notion of p-homogeneity.

Definition 2.1.2: p-Homogeneity

Consider topological spaces X and Y. A mapping $f\colon X \to Y$ is called **p-homogeneous** for some $p \in \mathbb{N}$ if it is positively homogeneous of degree p, i.e., for all $\lambda > 0$ and $x \in X$ we have $f(\lambda x) = \lambda^p f(x)$.

For some Banach space Q, a dissipation potential $\mathscr{R}\colon Q \times Q \to [0,\infty]$ is called **p-homogeneous** if the mapping $\mathscr{R}(q,\cdot)\colon Q \to [0,\infty]$ is p-homogeneous for every $q \in Q$.

Using this definition together with the one of the convex subdifferential from (1.3.2), we can directly identify an equivalent characterization of rate-independence for $(Q,\mathcal{E},\mathscr{R})$. While the statement itself is given in [66], we supplement it with the "trivial" proof lacking there:

Proposition 2.1.3: Rate-Independence via 1-Homogeneity

Consider a Banach space Q, a time interval $[0,T]$, an energy functional $\mathcal{E}\colon [0,T]\times Q \to \mathbb{R}$ and a general dissipation potential $\mathscr{R}\colon Q \times Q \to [0,\infty]$ as in (2.1.4).

Then, 1-homogeneity of $\mathscr{R}(q,\cdot)$ for all $q \in Q$ is equivalent to $(Q,\mathcal{E},\mathscr{R})$ being a rate-independent system.

Proof. In order to show rate-independence, we consider a solution $q \in W^{1,1}\big(0,T;Q\big)$ together with an arbitrary strictly increasing continuous time reparameterization $\alpha\colon [t_1,t_2] \to [t_1^*,t_2^*] \in \mathcal{J}_{[0,T]}$ with $\alpha(t_i) = t_i^*$, $i = 1,2$. Note that every monotone function is differentiable almost everywhere on its domain by a standard result first formulated by Lebesgue, cf. [35]. Apparently, where it exists, this derivative $\dot\alpha(t)$ is strictly positive.

Additionally, it is easy to see that an intuitive chain rule also holds for Banach space valued $W^{1,p}$-functions, which leads us to the fact that for the rescaled solution $\tilde q\colon [0,T] \to Q$, $\tilde q(t) \coloneqq q\big(\alpha(t)\big)$, we obtain $\frac{\mathrm{d}}{\mathrm{d}t}\tilde q(t) = \dot\alpha(t)\dot q\big(\alpha(t)\big)$, at least for almost every $t \in [t_1,t_2]$. This lets us conclude that rescaling of time for a solution as in Definition 2.1.1 simply scales the derivative by the corresponding factor.

Thus, verifying rate-independence is equivalent to showing that the subdifferential mapping of our dissipation functional is 0-homogeneous, i.e., we can write

$$\partial^v\mathscr{R}(q,\lambda v) = \lambda^0\partial^v\mathscr{R}(q,v) = \partial^v\mathscr{R}(q,v)$$

for any $\lambda > 0$ and $q,v \in Q$. To this end, we write the determining inequality for the subdifferential from (1.3.2) in the form

$$\forall \tilde v \in Q\colon \mathscr{R}(q,\tilde v) - \langle \xi,\tilde v\rangle \geq \mathscr{R}(q,v) - \langle \xi, v\rangle$$

and recognize that 0-homogeneity of $\partial^v\mathscr{R}(q,\cdot)$ is equivalent to 1-homogeneity of $\mathscr{R}(q,\cdot)$ by the linearity of $\xi \in Q^*$. This concludes the proof of the proposition. □

[1] For the sake of simplicity, we write $(Q,\mathcal{E},\mathscr{R})$ instead of more accurately $\big(Q,\mathcal{E},\mathscr{R}(q,\cdot)\big)$.

2.1.2 Problem Formulations and Solution Concepts

In what follows, we will deduce the solution concept which we will later on also pursue for our formulation and solution of problems in finite strain plasticity. We still mainly reproduce theory from [66] but give our own concise version of it here. To this end, we remember the differential inclusion

$$\partial^v \mathscr{R}\big(q(t), \dot{q}(t)\big) + \mathrm{D}_q \mathcal{E}\big(t, q(t)\big) \ni 0 \tag{2.1.6}$$

for some general dissipation potential as in (2.1.4) and an energy functional $\mathcal{E}\colon [0,T] \times Q \to \mathbb{R}$ on some time interval $[0,T]$. For this formulation of the equation driving our rate-independent system, we have already implicitly introduced a solution concept in Definition 2.1.1 which was rather focused on the system than on the time-dependent state solving (2.1.6). To this end, we formulate the following:

Definition 2.1.4: Differential Solutions

A mapping $q \in W^{1,1}\big([0,T];Q\big)$ is called a **differential solution** of the rate-independent system $(Q, \mathcal{E}, \mathscr{R})$ if the differential inclusion (2.1.6) holds in Q^* for almost all $t \in [0,T]$ together with $q(0) = q_0$ for some suitable initial condition $q_0 \in Q$.

In order to from here motivate and deduce the concept of so-called *energetic solutions*, we will first take a look at equivalent formulations of the inclusion problem (2.1.6) in the form of $-\mathrm{D}_q \mathcal{E}\big(t, q(t)\big) \in \partial^v \mathscr{R}\big(q(t), \dot{q}(t)\big)$. We can reformulate the latter directly via the characterizing inequality for the convex subdifferential and thereby obtain the so-called *evolutionary variational inequality*

$$\langle \mathrm{D}_q \mathcal{E}\big(t, q(t)\big), \tilde{v} - \dot{q}(t)\rangle + \mathscr{R}\big(q(t), \tilde{v}\big) - \mathscr{R}\big(q(t), \dot{q}(t)\big) \geq 0 \tag{2.1.7}$$

for all $\tilde{v} \in Q$ and almost all $t \in [0,T]$. For a physical interpretation, we transform (2.1.7) to

$$\mathscr{R}\big(q(t), \tilde{v}\big) - \mathscr{R}\big(q(t), \dot{q}(t)\big) \geq \langle -\mathrm{D}_q \mathcal{E}\big(t, q(t)\big), \tilde{v} - \dot{q}(t)\rangle$$

which equivalently states the force balance from (2.1.6) in Q^* insofar that, for any other rate of state-change $\tilde{v} \in Q$, the difference in the dissipation potential is larger (or equal) than the potential restoring force $-\mathrm{D}_q \mathcal{E}\big(t, q(t)\big)$ evaluated along the "difference vector" $\tilde{v} - \dot{q}(t)$.

Note that here we only took advantage of the definition of the convex subdifferential which implies that (2.1.7) holds for general (convex) dissipation potentials. The notion of rate-independence from Definition 2.1.1 does not appear in this formulation. As we have seen in Proposition 2.1.3, rate-independence enters our system directly via 1-homogeneity of the general dissipation potential in the form of $\mathscr{R}(q,\cdot)\colon Q \to [0,\infty]$ for any $q \in Q$. Since the general dissipation potential occurs via its subdifferential in (2.1.6), the investigation of convex subdifferentials of 1-homogeneous functionals is an intuitive next step. The following assertion together with its proof can be found in [66, Lemma 1.3.1]:

Lemma 2.1.5: Convex Subdifferentials of 1-Homogeneous Functionals

Let $\mathcal{R}\colon Q \to [0,\infty]$ be lower semi-continuous, convex, and 1-homogeneous. Then, we can characterize the convex subdifferential at any $v \in Q$ as

$$\partial \mathcal{R}(v) = \big\{ \eta \in \Sigma \,\big|\, \mathcal{R}(v) = \langle \eta, v\rangle \big\}\,, \quad \text{where } \Sigma := \partial \mathcal{R}(0)\,. \tag{2.1.8}$$

Moreover, we have the characterization

$$\xi \in \partial\mathcal{R}(v) \;\Leftrightarrow\; \forall w \in Q\colon \mathcal{R}(w) - \langle \xi, w\rangle \geq \mathcal{R}(v) - \langle \xi, v\rangle = 0\,.$$

Now, we will take the above characterization into account in order to reformulate the differential inclusion (2.1.6). As we can see in (2.1.8), the inclusion of some dual space element $\eta \in Q^*$ in the subdifferential $\partial\mathcal{R}(v)$ is split up into two separate statements. Firstly, the inclusion $\eta \in \Sigma \coloneqq \partial\mathcal{R}(0)$ where the latter set is often referred to as the *abstract elasticity domain* in the context. Secondly, the equality $\mathcal{R}(v) = \langle \eta, v\rangle$ has to hold.

In our case, where we have $\mathcal{R} = \mathscr{R}\big(q(t), \cdot\big)$ and $\eta = -\mathrm{D}_q\mathcal{E}\big(t, q(t)\big)$, this split also leads to two separate relations governing the rate-independent system. These can be easily reformulated as the following conditions which have to hold for almost all $t \in [0,T]$:

local stability:	$\forall v \in Q\colon \langle \mathrm{D}_q\mathcal{E}\big(t, q(t)\big), v\rangle + \mathscr{R}\big(q(t), v\big) \geq 0$	(2.1.9a)
power balance:	$\langle \mathrm{D}_q\mathcal{E}\big(t, q(t)\big), \dot{q}(t)\rangle + \mathscr{R}\big(q(t), \dot{q}(t)\big) = 0$	(2.1.9b)

Let us also here comment on the physical interpretation of this split of conditions describing the system under consideration: The local stability statement (2.1.9a), which stems from $0 \in \partial^v\mathscr{R}\big(q(t), 0\big) + \mathrm{D}_q\mathcal{E}\big(t, q(t)\big)$, is a purely static condition since it does not include any time derivatives $\dot{q}(t)$. Physically, this condition says that already the static frictional forces $\partial^v\mathscr{R}\big(q(t), 0\big)$ must be strong enough to balance the potential restoring force $-\mathrm{D}_q\mathcal{E}\big(t, q(t)\big)$. Again, this is a force balance in Q^* which becomes apparent by the fact that (2.1.9a) holds for all other possible states $v \in Q$. Invariance under time-rescaling is apparent here since no time-derivatives are incorporated.

The second condition (2.1.9b), which completes the description of our rate-independent system, is the power balance which considers the power of the state-change $\langle -\mathrm{D}_q\mathcal{E}\big(t, q(t)\big), \dot{q}(t)\rangle$, i.e., the potential restoring force from before evaluated along the tangential state-vector in time. This has to be equal to the dissipation rate $\mathscr{R}\big(q(t), \dot{q}(t)\big)$. Here, the 1-homogeneity of $\mathscr{R}\big(q(t), \cdot\big)$ yields that the solution property of rescaled solutions remains preserved.

All in all, we can summarize the evolution of the rate-independent system as the purely static condition (2.1.9a) and the scalar power relation (2.1.9b) which incorporates change rates and thereby establishes a connection between state equations for fixed $t \in [0,T]$. As mentioned above, both conditions behave reasonably when exposed to time-rescaling of solutions and thus also preserve our understanding of rate-independence.

The Maximum Dissipation Principle

Even though our ultimate goal for this section is the motivation and definition of energetic solutions for rate-independent systems, we will now devote some short deliberations to the physically motivated notion of the *principle of maximal dissipation*. As we will see later on in Section 2.2.2, this relation is very illustrative in the context of motivating finite strain plasticity which is why we will attribute some importance also to its formulation in general rate-independent systems.

The derivation of the maximum dissipation principle can be summarized as a reformulation of (2.1.9b) by results of convex analysis. To this end, we define the *convex conjugate*[2] of our

[2]Convex conjugates are sometimes also referred to as *Legendre-Fenchel Transforms*.

general dissipation potential by

$$\big[\mathscr{R}(q,\cdot)\big]^*\colon Q^* \to]-\infty,\infty] \quad , \quad \big[\mathscr{R}(q,\cdot)\big]^*(\xi) \coloneqq \sup_{v\in Q} \langle \xi, v\rangle - \mathscr{R}(q,v) \tag{2.1.10}$$

for any $q \in Q$. A thorough introduction to the concept of convex conjugates and their use in convex analysis can be found in classical works like [25, Chapter I, Section 4] or [90, Chapter 11]. Rate-independence of our system yields that the following representation of the convex conjugate of the dissipation potential holds:

Lemma 2.1.6: Convex Conjugate of the Dissipation Potential

The convex conjugate (2.1.10) of the 1-homogeneous $\mathscr{R}(q,\cdot)\colon Q \to \mathbb{R}$ takes the form

$$\big[\mathscr{R}(q,\cdot)\big]^*(\xi) = \mathcal{X}_{\Sigma(q)}(\xi) \coloneqq \begin{cases} 0 & , \text{ if } \xi \in \Sigma(q) \\ \infty & , \text{ otherwise} \end{cases} \tag{2.1.11}$$

where the latter function is referred to as the **characteristic function** $\mathcal{X}_{\Sigma(q)}$ of the **abstract elasticity domain** $\Sigma(q) \coloneqq \partial^v \mathscr{R}(q,0)$ for any state $q \in Q$. Thus, we have the identity

$$\mathscr{R}(q,v) = \max_{\tilde{\xi}\in\Sigma(q)} \langle \tilde{\xi}, v\rangle \quad \text{for any} \quad q, v \in Q\,. \tag{2.1.12}$$

Proof. The 1-homogeneity of $\mathscr{R}(q,\cdot)$ as in Definition 2.1.2 directly implies $\mathscr{R}(q,0) = 0$ for any $q \in Q$. Now, consider $q \in Q$ fixed. For some arbitrary $\xi \in Q^*$, we now distinct the cases of inclusion into $\Sigma(q) = \partial^v \mathscr{R}(q,0)$:

In the case of $\xi \in \partial^v \mathscr{R}(q,0)$, the subdifferential definition (1.3.2) immediately gives

$$0 \geq \langle \xi, \tilde{v}\rangle - \mathscr{R}(q,\tilde{v}) \quad \text{for any} \quad \tilde{v} \in Q$$

where the right-hand side in particular is equal to zero for $\tilde{v} = 0$. Thus, we also have

$$\big[\mathscr{R}(q,\cdot)\big]^*(\xi) = \sup_{v\in Q} \langle \xi, v\rangle - \mathscr{R}(q,v) = 0$$

for $\xi \in \partial^v \mathscr{R}(q,0)$. In the remaining case we have $0 < \langle \xi, \bar{v}\rangle - \mathscr{R}(q,\bar{v})$ for some $\bar{v} \in Q$ where the right-hand side can be scaled arbitrarily by the 1-homogeneity of $\mathscr{R}(q,\cdot)$ and the linearity of $\xi \in Q^*$. Thus, it is not bounded from above, the respective supremum from definition (2.1.10) does not exist in $\mathbb{R}$ and $\big[\mathscr{R}(q,\cdot)\big]^*(\xi) = \infty$ holds. This concludes the proof of (2.1.11).

From here, (2.1.12) is in direct reach. Due to the convexity of $\mathscr{R}(q,\cdot)$, we can take advantage of the biduality of proper, l.s.c. and convex functions, cf. [90, Theorem 11.1], together with the definition of the convex conjugate and the indicator function $\mathcal{X}_{\Sigma(q)}$ in order to obtain

$$\mathscr{R}(q,v) = \big[\mathcal{X}_{\Sigma(q)}\big]^*(v) = \sup_{\tilde{\xi}\in Q^*} \langle \tilde{\xi}, v\rangle - \mathcal{X}_{\Sigma(q)}(\tilde{\xi}) = \sup_{\tilde{\xi}\in\Sigma(q)} \langle \tilde{\xi}, v\rangle$$

for any $q, v \in Q$. The particular structure of the abstract elasticity domain yields that the latter supremum is attained and can thus be written as a maximum. □

With the aid of (2.1.12), we can now reformulate the power balance (2.1.9b) via

$$\langle \xi(t), \dot{q}(t) \rangle = \max_{\tilde{\xi} \in \Sigma(q(t))} \langle \tilde{\xi}, \dot{q}(t) \rangle \text{ with } \xi(t) \in -\partial^q \mathcal{E}\big(t, q(t)\big) \text{ and } \Sigma(q) \coloneqq \partial^v \mathscr{R}(q, 0) \tag{2.1.13}$$

where we also generalized the setting to possibly non-smooth, convex energy functionals $\mathcal{E}(t, \cdot)$. In the smooth case, we obviously have $\xi(t) = \mathrm{D}_q \mathcal{E}(t, q)$. The adaption allows us to distinguish between the not necessarily unique *actual driving force* ξ and the set of *available driving forces* $-\partial^q \mathcal{E}(t, q)$.

The *maximum dissipation principle* as formulated in (2.1.13) now states that the actual driving force maximizes the dissipation for fixed order-parameter rate $\dot{q}(t)$. The possible driving force $\tilde{\xi} \in Q^*$ varies freely over the set of all admissible driving forces $\Sigma\big(q(t)\big)$ which coincides with the abstract elasticity domain introduced beforehand.

Another interpretation of (2.1.13) implies that the rate $\tilde{q}(t)$ is orthogonal to the abstract elasticity domain, formerly introduced as the *orthogonality principle* in [118] which further generalizes *Onsager's principle* from [77]. As a consequence, the state q cannot evolve if the driving force is in the interior of $\Sigma(q)$. This understanding is compatible with the illustration and development of plastic behavior described in Section 2.2.2 but for general rate-independent systems provides a rather theoretical point of view.

As the previous arguments have demonstrated, the maximum dissipation principle (2.1.13) is equivalent to the power balance (2.1.9b) and thus, if augmented with the local stability condition (2.1.9a), fully describes the evolution in time of our rate-independent system according to the inclusion formulation (2.1.6).

Generalization to Energetic Solutions

All of the previous problem formulations and solution concepts intrinsically require differentiability of respective solutions in time. Generally, we cannot even expect absolute continuity of solutions with respect to time but have to include solutions with jumps into our framework. Additionally, for our application of the theory to finite strain elasto-plasticity later on, we cannot even assume the state space Q to have linear structure since the corresponding problem is formulated on a manifold. For this reason, we will avoid derivatives with respect to time and state but still stick to the structure from above, i.e., have a static stability condition and an energy inequality which we often formulate directly as an energy balance.

For this reason, the next step on the road to the definition of energetic solutions is an integrated form of the power balance (2.1.9b). Thus, we take advantage of the chain rule for the total time derivative

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}\big(s, q(s)\big) = \langle \mathrm{D}_q \mathcal{E}\big(s, q(s)\big), \dot{q}(s) \rangle + \mathrm{D}_t \mathcal{E}\big(s, q(s)\big) \quad \text{for all } s \in [0, T]$$

together with the fundamental theorem of calculus in order to obtain

$$\mathcal{E}\big(t, q(t)\big) + \int_r^t \mathscr{R}\big(q(s), \dot{q}(s)\big) \mathrm{d}s = \mathcal{E}\big(r, q(r)\big) + \int_r^t \mathrm{D}_t \mathcal{E}\big(s, q(s)\big) \mathrm{d}s \,. \tag{2.1.14}$$

We can interpret the partial time derivative $\mathrm{D}_t \mathcal{E}(s, q(s))$ as an actual power induced by the temporal changes in the system. Since (2.1.9b) had to be fulfilled for almost all $t \in [0, T]$ the integrated identity (2.1.14) now holds for all $r, t \in [0, T]$.

Unfortunately, the integrated formulation above still contains a time derivative $\dot{q}$ of our solution. In order to remedy this flaw, we will generalize (2.1.14) even further. To this end, we observe that the integral over $\mathscr{R}(q,\dot{q})$ measures the total dissipation along the curve q (which is our solution) between times r and t. This total dissipation term can also be modelled in a more general sense as we introduce the *dissipation distance* $\mathscr{D}\colon Q \times Q \to [0,\infty]$ which canonically can be conceived to be of the form $\mathscr{D}(q_1,q_2) = \mathcal{R}(q_2 - q_1)$ for some dissipation functional $\mathcal{R}$ and states $q_1, q_2 \in Q$. In general, we assume $\mathscr{D}$ to satisfy the triangle inequality but allow it to be asymmetric.

Thus, $\mathscr{D}(q_1,q_2)$ measures the minimal amount of energy which is dissipated when the state changes from q_1 to q_2. With this thought in mind, we can now formulate a new notion of the *total dissipation* along a part $[r,s] \subset [0,T]$ of a parameterized curve $q\colon [0,T] \to Q$ via

$$\mathrm{Diss}_{\mathscr{D}}\big(q;[r,s]\big) \coloneqq \sup\left\{ \sum_{j=1}^{N} \mathscr{D}\big(q(t_{j-1}),q(t_j)\big) \,\middle|\, N \in \mathbb{N},\, r \le t_0 < t_1 < \ldots < t_{N-1} < t_N \le s \right\}.$$

This notion of dissipation along curves can also be interpreted as the total variation with respect to the dissipation distance $\mathscr{D}$. It now lets us transform the power balance (2.1.9b) via the integrated formulation (2.1.14) into an *energy balance*

$$\mathcal{E}\big(t,q(t)\big) + \mathrm{Diss}_{\mathscr{D}}\big(q;[0,t]\big) = \mathcal{E}\big(0,q(0)\big) + \int_0^t \mathrm{D}_t\mathcal{E}\big(s,q(s)\big)\mathrm{d}s \tag{2.1.15}$$

for every $t \in [0,T]$ and some adequately defined dissipation distance $\mathscr{D}$.

The last conceptual generalization which we introduce before finally turning our attention to the definition of energetic solutions concerns the local stability condition (2.1.9a). This formulation still contains the differential $\mathrm{D}_q\mathcal{E}$ which demands a linear structure of the underlying state space Q. Whenever the corresponding energy functional $\mathcal{E}(t,\cdot)$ is convex, (2.1.9a) is equivalent to the *global stability condition*

$$\mathcal{E}\big(t,q(t)\big) \le \mathcal{E}\big(t,\tilde{q}\big) + \mathscr{R}\big(q(t),\tilde{q}-q(t)\big) \quad \text{for all } \tilde{q} \in Q\,. \tag{2.1.16}$$

While convexity of the energy directly yields equivalence between (2.1.9a) and (2.1.16), for general energy functionals the global stability merely implies the local one but not the other way around.

The newly achieved absence of derivatives now also allows us to omit the linear structure of our Banach space Q and thus replace it with some general space $\mathcal{Q}$. Additionally, the above formulation of the energy balance (2.1.15) does not depend on a dissipation potential $\mathscr{R}$ as before but only relies on the dissipation distance $\mathscr{D}$. These adaptations of the formulation for rate-independent systems $(Q,\mathcal{E},\mathscr{R})$ from Definition 2.1.1 lead us to the new notion of *Energetic Rate-Independent Systems* (ERIS) determined by the triple $(\mathcal{Q},\mathcal{E},\mathscr{D})$ for the description of our problem at hand.

Definition 2.1.7: Energetic Rate-Independent Systems and Solutions

A function $q\colon [0,T] \to \mathcal{Q}$ is called an **energetic solution** of the ERIS $(\mathcal{Q},\mathcal{E},\mathscr{D})$ if it

satisfies the **stability condition** (S) and the **energy balance** (E) for all $t \in [0,T]$:

$$\forall\, \tilde{q} \in \mathcal{Q}\colon \quad \mathcal{E}\big(t, q(t)\big) \leq \mathcal{E}\big(t, \tilde{q}\big) + \mathscr{D}\big(q(t), \tilde{q}\big)\,, \tag{S}$$

$$\mathcal{E}\big(t, q(t)\big) + \mathrm{Diss}_{\mathscr{D}}\big(q; [0,t]\big) = \mathcal{E}\big(0, q(0)\big) + \int_0^t \mathrm{D}_t\mathcal{E}\big(s, q(s)\big)\mathrm{d}s\,. \tag{E}$$

The rate-independence as characterized in Definition 2.1.1 of the energetic formulation via (S) and (E) from Definition 2.1.7 is apparent due to the purely static nature of the global stability condition (S) and the fact that the energy balance (E) is also invariant under time rescaling of solutions. Thus, referring also to ERIS $(\mathcal{Q}, \mathcal{E}, \mathscr{D})$ as rate-independent systems is justified. Augmenting the above definition with a suitable initial condition $q(0) = q_0 \in \mathcal{Q}$, we obtain the initial value problem $(\mathcal{Q}, \mathcal{E}, \mathscr{D}, q_0)$.

Formulating an initial value problem for a notion of solutions which does not require solutions to be absolutely continuous in time and allows for jumps might sound controversial at first and the question arises, in which way the initial value is actually attained. In (E), $q(0)$ appears which allows us to demand this estimate with the required initial value $q_0 \in \mathcal{Q}$. Additionally, we will see later that the way in which we construct solutions via time-incremental problems also allows to prescribe initial states of the solutions in a natural way.

In addition to the thoroughly discussed absence of derivatives and the resulting generalization to spaces $\mathcal{Q}$, we will later on benefit from the energetic formulation in Definition 2.1.7 insofar that the corresponding existence theory can be derived solely from suitable lower semicontinuity properties for $\mathcal{E}$ and $\mathscr{D}$ together with some compactness assumptions. We will construct solutions via so-called *time-incremental minimization problems* for which we consider the *set of all partitions* of some interval $[r,s]$ given by

$$\mathrm{Part}\big([r,s]\big) \coloneqq \big\{(t_0, ..., t_N) \,\big|\, r = t_0 < t_1 < ... < t_N = s\big\}\,.$$

Important quantities of each partition $\Pi \in \mathrm{Part}\big([r,s]\big)$ include its *number of subintervals* $N_\Pi \coloneqq N$ from the above definition and its *fineness* $\varnothing(\Pi) \coloneqq \max\{t_k - t_{k-1} \,|\, k = 1, ..., N_\Pi\}$.

Given an initial condition $q_0 \in \mathcal{Q}$ and partition $\Pi \in \mathrm{Part}\big([0,T]\big)$, we can thus formulate

$$(\mathrm{IMP})^\Pi \qquad q_k \in \operatorname*{arg\,min}_{\tilde{q} \in \mathcal{Q}} \mathcal{E}\big(t_k, \tilde{q}\big) + \mathscr{D}\big(q_{k-1}, \tilde{q}\big) \quad \text{for } k = 1, ..., N_\Pi\,.$$

We will use the solutions of these incremental problems, i.e., the minimizers of $\mathcal{E}\big(t_k, \tilde{q}\big) + \mathscr{D}\big(q_{k-1}, \tilde{q}\big)$, in order to define sequences of piecewise constant interpolants which will – under sufficient assumptions – converge to an energetic solution of the corresponding ERIS for $\varnothing(\Pi) \to 0$.

These incremental problems will constitute the underlying composite minimization problems for the algorithmic deliberations of Chapters 3 and 4 and are thus of crucial importance not only for the analysis of the problem class but also for the application of algorithms developed for it later on.

2.1.3 Existence of Energetic Solutions

Before we now recall sufficient conditions from [66] under which energetic solutions defined via Definition 2.1.7 exist, we will introduce an important modification to the above system formulation. To this end, we split the state space $\mathcal{Q}$ into a non-dissipative component $\mathcal{Y}$ and

a dissipative part $\mathcal{Z}$ via $\mathcal{Q} = \mathcal{Y} \times \mathcal{Z}$. As the denotation suggests, we thus assume that the dissipation potential $\mathscr{R}$ only depends on the z-component in form of

$$\mathscr{R}(q, \dot{q}) = \mathscr{R}(z, \dot{z}) \quad \text{and} \quad \big(\mathscr{R}(z, \dot{z}) = 0 \;\Rightarrow\; \dot{z} = 0\big)\,.$$

In that case, the differential inclusion (2.1.2) can be reformulated as the coupled system

$$\mathrm{D}_y \mathcal{E}(t, y, z) = 0 \quad \text{and} \quad 0 \in \partial^v \mathscr{R}(z, \dot{z}) + \mathrm{D}_z \mathcal{E}(t, y, z)\,. \tag{2.1.17}$$

Here, we can see that the split in components also leads to a different treatment of the y- and the z-component. In particular, the so-called reduced energy functional

$$\mathcal{J}(t, z) \coloneqq \min_{y \in \mathcal{Y}} \mathcal{E}(t, y, z) \tag{2.1.18}$$

automatically suffices the first equation of (2.1.17) such that we are left with the differential inclusion

$$0 \in \partial^v \mathscr{R}(z, \dot{z}) + \mathrm{D}_z \mathcal{J}(t, z) \tag{2.1.19}$$

and thus call $\big(\mathcal{Z}, \mathcal{J}, \mathscr{R}\big)$ the *reduced RIS*. This reformulation of rate-independent systems is also reversible insofar that once $z\colon [0,T] \to \mathcal{Z}$ solves (2.1.19), we can recover the non-dissipative component by determining $y(t) \coloneqq \arg\min_{y \in \mathcal{Y}} \mathcal{E}\big(t, y, z(t)\big)$ such that then $q(t) \coloneqq \big(y(t), z(t)\big)$ solves the original system.

From the standpoint of applications, we often (but not always) can view y as the observable variables while z refers to internal variables which are neither directly observable nor controllable from the outside. This interpretation of internal variables is emphasized by the lack of time derivatives $\dot{y}$ in (2.1.17) which implies that instantaneous changes of the observable variable y cannot influence the changes of z. Thus, often y is referred to as the "fast" component while the internal variable z is called "slow" in the context.

Assumptions on the Dissipation Distance and Stored Energy Functional

Let us now start with the formulation of sufficient assumptions for the existence of energetic solutions. The first topic to be handled is the dissipation distance $\mathscr{D}\colon \mathcal{Z} \times \mathcal{Z} \to [0, \infty]$, which naturally only depends on the dissipative component and is assumed to be an *extended quasi-distance*, i.e., we demand

$$\begin{aligned} &(i) \quad \forall z_1, z_2, z_3 \in \mathcal{Z}\colon \mathscr{D}(z_1, z_3) \le \mathscr{D}(z_1, z_2) + \mathscr{D}(z_2, z_3) \\ &(ii) \quad \forall z_1, z_2 \in \mathcal{Z}\colon \mathscr{D}(z_1, z_2) = 0 \Leftrightarrow z_1 = z_2 \end{aligned} \tag{D1}$$

allowing both non-symmetry and infinite values opposed to traditional distance measures. Additionally, we require that

$$\mathscr{D}\colon \mathcal{Z} \times \mathcal{Z} \to [0, \infty] \quad \text{is lower-semicontinuous.} \tag{D2}$$

As far as the stored energy functional $\mathcal{E}\colon [0,T] \times \mathcal{Q} \to \mathbb{R} \cup \{\infty\}$ is concerned, the assumptions are a bit more involved. The first one is the *compactness of sublevels*:

$$\forall\, t \in [0,T]\colon\ \mathcal{E}(t, \cdot)\colon \mathcal{Q} \to]-\infty, \infty] \quad \text{has compact sublevels.} \tag{E1}$$

The second one concerns the *domain* $\operatorname{dom}\mathcal{E} \coloneqq \big\{(t,q) \in [0,T] \times \mathcal{Q} \;\big|\; \mathcal{E}(t,q) < \infty\big\}$ as well as for any fixed $t \in [0,T]$ correspondingly $\operatorname{dom}\mathcal{E}(t,\cdot) \coloneqq \big\{q \in \mathcal{Q} \;\big|\; \mathcal{E}(t,q) < \infty\big\}$. Thus, we formulate the *energetic control of power* given by:

$$\begin{aligned}
&\operatorname{dom}\mathcal{E} = [0,T] \times \operatorname{dom}\mathcal{E}(t,\cdot)\,,\\
&\exists\, c_{\mathcal{E}} \in \mathbb{R}, \lambda_{\mathcal{E}} \in L^1(0,T), N_{\mathcal{E}} \subset [0,T] \text{ of measure zero}\\
&\forall\, q \in \operatorname{dom}\mathcal{E}(0,\cdot)\colon \mathcal{E}(\cdot,q) \in W^{1,1}(0,T)\,, \qquad\qquad \text{(E2)}\\
&\qquad\qquad \mathrm{D}_t\mathcal{E}(t,q) \text{ exists for } t \in [0,T] \setminus N_{\mathcal{E}} \text{ and satisfies}\\
&\qquad\qquad |\mathrm{D}_t\mathcal{E}(t,q)| \leq \lambda_{\mathcal{E}}(t)\big(\mathcal{E}(t,q) + c_{\mathcal{E}}\big)\,.
\end{aligned}$$

While this assumption might seem rather opaque at first sight, it simply ensures sufficient regularity of the stored energy functional in time together with a bound which enables the application of Gronwall's inequality. The latter lets us conclude that the sublevels of $\mathcal{E}$ as a whole are compact if and only if the sublevels of $\mathcal{E}(t,\cdot)$ are compact for all $t \in [0,T]$ which is given by the prior assumption (E1). Compactness of sublevels on the other hand will be very useful as far as the convergence of adequately chosen subsequences of approximate solutions is concerned. Additionally, the compactness of non-empty sublevels of $\mathcal{E}(t,\cdot)$ yields lower semi-continuity of said mapping which is important for solvability of the time-incremental problems from $(\mathrm{IMP})^{\Pi}$.

The stability condition (S) can also be formulated via time-dependent so-called *sets of stable states* at some time $t \in [0,T]$ which are defined via

$$S(t) \coloneqq \big\{q = (y,z) \in \mathcal{Q} \;\big|\; \mathcal{E}(t,q) < \infty,\ \forall \hat{q} = (\hat{y},\hat{z}) \in \mathcal{Q}\colon \mathcal{E}(t,q) \leq \mathcal{E}(t,\hat{q}) + \mathscr{D}(z,\hat{z})\big\}$$

such that (S) simply means that $q(t) \in S(t)$ holds for all $t \in [0,T]$. The properties of these sets of stable states turn out to be crucial for deriving existence results.

With the assumptions on $\mathcal{E}$ and $\mathscr{D}$ motivated as well as the reformulation of the stability condition at hand, key properties of the solutions of time-incremental problems $(\mathrm{IMP})^{\Pi}$ can be verified, cf. [66, Proposition 2.1.4]:

Proposition 2.1.8: Properties of Time-Incremental Solutions

Let (D1) and (E2) hold. Every solution of $(\mathrm{IMP})^{\Pi}$ satisfies the following properties:

(i) $\forall k \in \{1,\dots,N_{\Pi}\}\colon q_k \in S(t_k)$, i.e., q_k is stable at time t_k.

(ii) $\forall k \in \{1,\dots,N_{\Pi}\}\colon \int_{t_{k-1}}^{t_k} \mathrm{D}_t\mathcal{E}(s,q_k)\,\mathrm{d}s \leq e_k - e_{k-1} + \delta_k \leq \int_{t_{k-1}}^{t_k} \mathrm{D}_t\mathcal{E}(s,q_{k-1})\,\mathrm{d}s$
where we denoted $e_j \coloneqq \mathcal{E}(t_j,q_j)$ and $\delta_k \coloneqq \mathscr{D}(z_{k-1},z_k)$.

(iii) If additionally (D2) and (E1) hold, then solutions of $(\mathrm{IMP})^{\Pi}$ exist.

In order to now transform these time-incremental solutions, which are only defined in discrete time points t_k, into functions actually approximating energetic solutions, piecewise constant interpolants of the q_k have to be defined. For these interpolants, crucial a priori bounds can be verified. These a priori bounds can then be taken advantage of in the following way: We choose a sequence of partitions Π_j of the time-interval $[0,T]$ the fineness $\varnothing\big(\Pi_j\big)$ of which converges to zero, and consider the corresponding solutions of the time-incremental problems $(\mathrm{IMP})^{\Pi_j}$. From this sequence of solutions, the formerly deduced bounds allow us to

extract a converging subsequence by a suitable version of Helly's selection principle.[3] This is at least possible for the z-component of said solutions for which we can control oscillations in time better with the aid of the dissipation terms within above estimates.

For the non-dissipative y-component, however, we have to be a bit more deliberate about the choice of converging subsequences. One possibility would be to additionally choose subsequences for every $t \in [0,T]$ and thereby rely on the axiom of choice for our argumentation (as in e.g. [30, 61]). We, however, keep following the steps of the authors in [66] who take another path and assume metrizability of the underlying topology[4] which is a slightly stronger assumption in itself but yields a simpler convergence proof and guarantees the existence of solutions which are measurable in time.

Compatibility Conditions and Existence of Energetic Solutions

As we have now picked converging subsequences and identified the corresponding limit functions, we need to still make sure that the latter suffice the properties demanded of energetic solutions according to Definition 2.1.7. To this end, we need to formulate further assumptions that involve both $\mathcal{E}$ and $\mathscr{D}$ expressing their compatibility since our prior assumptions only consider either the stored energy or the dissipation on their own. The definition of these conditions requires the notion of *stable sequences* $(t_m, q_m)_{m\in\mathbb{N}}$ given by

$$\sup_{m\in\mathbb{N}} \mathcal{E}(t_m, q_m) < \infty \quad \text{and} \quad \forall m \in \mathbb{N}\colon\ q_m \in S(t_m)\,.$$

The boundedness condition here intrinsically links the type of convergence of states to the properties of $\mathcal{E}$ and $\mathscr{D}$. Convergent stable sequences then help us to characterize the *compatibility* of these functionals via

$$\begin{aligned}
&\forall \text{ stable sequences } (t_m, q_m)_{m\in\mathbb{N}} \text{ with } (t_m, q_m) \overset{[0,T]\times\mathcal{Q}}{\longrightarrow} (t,q)\,:\\
&\quad t \in [0,T]\setminus N_{\mathcal{E}} \text{ with } N_{\mathcal{E}} \text{ from (E2)} \ \Rightarrow\ \mathrm{D}_t\mathcal{E}(t,q) = \lim_{m\to\infty} \mathrm{D}_t\mathcal{E}(t,q_m)\,, && \text{(C1)}\\
&\quad q \in S(t)\,. && \text{(C2)}
\end{aligned}$$

The first condition (C1) is referred to as the *conditioned continuity of the power of external forces* and the second one (C2) is intuitively called *closedness of the stability set*. We formulated the conditions just in the way they are utilized within the proof of existence of energetic solutions but verifying their validity is often achieved via alternative characterizations. We will not elaborate on the latter but refer to [66, Section 2.1.5] for further information on these reformulations.

Now that we have introduced sufficiently strong assumptions on the stored energy functional $\mathcal{E}$ and the dissipation functional $\mathscr{D}$ as well as have outlined their contributions to the proof of existence of energetic solutions, we can finally formulate the central result of this section as found in [66, Theorem 2.1.6]. While we have tried to at least sketch the proof within

[3] Named after Austrian mathematician Eduard Helly (1884-1943) – states that uniformly bounded sequences of monotone real valued functions include a convergent subsequence. It can also be generalized to the compactness of the space BV_{loc} of functions of locally bounded variation, cf. [74, VIII.§4].

[4] Metrizability of a topological space $(X,\mathcal{T})$ describes the possibility to define a metric $d\colon X\times X \to [0,\infty[$ such that the topology induced by d is $\mathcal{T}$. A famous example for a metrizability criterion is Urysohn's Metrization Theorem (Pawel Urysohn, 1898-1924, cf. [113]) equivalent to a topological space being separable and metrizable if and only if it is regular.

the motivation of the above conditions, [66, Section 2.1.6] is entirely dedicated to giving a detailed and comprehensive version of the argumentation.

Theorem 2.1.9: Existence of Energetic Solutions

Assume that $\mathcal{E}$ and $\mathscr{D}$ satisfy the assumptions (D1)-(D2) and (E1)-(E2) together with the compatibility conditions (C1)-(C2). Furthermore, assume that the topology of $\mathcal{Q}$ restricted to compact sets is separable and metrizable. Then, the following assertions hold:

(i) For each $q_0 \in S(0)$, there exists an energetic solution $q = (y, z)\colon [0,T] \to \mathcal{Q}$ to the initial value problem $(\mathcal{Q}, \mathcal{E}, \mathscr{D}, q_0)$. Moreover, $q\colon [0,T] \to \mathcal{Q}$ is measurable and for almost all $t \in [0,T]$ we have:

$$\mathrm{D}_t\mathcal{E}\big(t, q(t)\big) = \mathrm{D}_t\mathcal{E}\big(t, y(t), z(t)\big) = \sup\left\{ \mathrm{D}_t\mathcal{E}(t,y,z) \,\middle|\, y \in \operatorname*{arg\,min}_{\tilde{y}\in\mathcal{Y}} \mathcal{E}(t,\tilde{y},z) \right\}.$$

(ii) If $\big(\Pi_l\big)_{l\in\mathbb{N}} \subset \mathrm{Part}\big([0,T]\big)$ is a sequence of partitions with fineness $\varnothing\big(\Pi_l\big) \to 0$ for $l \to \infty$ and $\underline{q}^{\Pi_l}$ is the interpolant of a solution of the associated $(\mathrm{IMP})^{\Pi_l}$, then there exist a subsequence $q_k \coloneqq \underline{q}^{\Pi_{l_k}}$ and an energetic solution $\tilde{q} = (\tilde{y}, \tilde{z})$ to the initial value problem $(\mathcal{Q}, \mathcal{E}, \mathscr{D}, q_0)$ such that the following holds:

$$\forall\, t \in [0,T]\colon \qquad z_k(t) \overset{\mathcal{Z}}{\to} \tilde{z}(t)\,, \tag{2.1.20a}$$
$$\forall\, t \in [0,T]\colon \qquad \mathrm{Diss}_{\mathscr{D}}\big(z_k; [0,t]\big) \to \mathrm{Diss}_{\mathscr{D}}\big(\tilde{z}; [0,t]\big)\,, \tag{2.1.20b}$$
$$\forall\, t \in [0,T]\colon \qquad \mathcal{E}\big(t, q_k(t)\big) \to \mathcal{E}\big(t, \tilde{q}(t)\big)\,, \tag{2.1.20c}$$
$$\forall_{a.a.}\, t \in [0,T]\colon \qquad \mathrm{D}_t\mathcal{E}\big(t, q_k(t)\big) \to \mathrm{D}_t\mathcal{E}\big(t, \tilde{q}(t)\big)\,. \tag{2.1.20d}$$

Moreover, (E2) and (2.1.20d) imply $\mathrm{D}_t\mathcal{E}\big(t, q_k(t)\big) \to \mathrm{D}_t\mathcal{E}\big(t, \tilde{q}(t)\big)$ in $L^1(0,T)$.

(iii) Suppose that the functional $\mathcal{E}(t,\cdot,z)$ has a unique minimizer y for each stable point $q \in (y,z) \in S(t)$. Then, taking $\tilde{y}(t) \coloneqq \operatorname{arg\,min} \mathcal{E}\big(t,\cdot,\tilde{z}(t)\big)$ improves the convergence in (2.1.20a) to

$$\forall\, t \in [0,T]\colon q_k(t) \overset{\mathcal{Q}}{\to} \tilde{q}(t)\,.$$

The above existence result provides the basis for our plans to approximate solutions of the finite strain plasticity problem numerically not only insofar that these energetic solutions do exist in the first place but also insofar that they can be approximated by converging sequences of time-incremental solutions. The question of exactly which form the corresponding time-incremental minimization problems $(\mathrm{IMP})^{\Pi}$ take will be answered in Section 2.2 and the question of how to then efficiently solve these problems numerically will be considered in Chapters 3 and 4.

2.2 Formulation of Finite Strain Plasticity

Now that we are aware of the theoretical framework which we want to place our finite strain plasticity application problem in, we still have to consider how the latter fits into it. To this end, in Section 2.2.1 we will give a continuum mechanical introduction to quasi-static hyperelasticity which constitutes the foundation of the underlying theory. Afterwards, we will build on that foundation by augmenting our until now purely elastic model with characteristic features of plasticity in Section 2.2.2. After formulating the latter problem of continuum mechanics as a rate-independent system in the sense of Section 2.1, we investigate sufficient assumptions on the corresponding functionals and domain spaces for the existence of energetic solutions in Section 2.2.3. Having then dealt with the problem in sufficient generality, in Section 2.2.4 we deduce the specific form of the application problem for which we will employ our solution algorithm in the later stages of the present treatise.

Notational Remark

Before departing on the endeavor mapped out above, let us give a notational remark: In order to distinguish between mappings on the $\mathbb{R}^3$-domain Ω (, or $\Omega \times [0,T]$ respectively,) and their placeholders within other functionals, we will always write the former in bold. As an example, we will have a matrix-valued mapping $\mathbf{P}\colon \overline{\Omega} \times [0,T] \to \mathrm{SL}(d)$ to be inserted into e.g. energy density functionals. In the definition of the latter, general matrix variables will then be referred to as P in non-bold. This strategy particularly reveals potential space- and time-dependencies even when they are mostly omitted for the sake of notational simplicity. All in all, we try to find a healthy balance between requiring common sense and pursuing notational accuracy.

2.2.1 Quasi-Static Hyperelasticity

The first goal is to motivate and deduce the formulation of the time-dependent deformation problem for finite strain elasticity with a constitutive equation governed by a stored energy functional. Therefore, we consider a body in undeformed shape at initial time $t = 0$ described by the set of points $\overline{\Omega}$ for an open, non-empty set $\Omega \in \mathbb{R}^d$ for $d \in \{2, 3\}$. The closure $\overline{\Omega}$ is called the *reference configuration* the boundary of which is split up to disjoint subsets $\Gamma_D, \Gamma_N \subset \partial\Omega$ for Dirichlet and Neumann conditions, respectively, such that $\overline{\Gamma}_D \cup \overline{\Gamma}_N = \partial\Omega$.

Apparently, the theory of elasticity evolves around the change of this reference configuration on account of internal and external forces acting on the body. The modelling of change within the material configuration is based on the *deformation vector field* $\boldsymbol{y}$ and the corresponding *deformed configuration* $\overline{\Omega^{t,\boldsymbol{y}}}$ at time $t \in [0,T]$ defined via

$$\boldsymbol{y}\colon \overline{\Omega} \times [0,T] \to \mathbb{R}^d \quad \text{and} \quad \overline{\Omega^{t,\boldsymbol{y}}} \coloneqq \boldsymbol{y}\big(\overline{\Omega}, t\big)$$

for which we will omit the time-dependence where not necessary in the context. For the mapping $\boldsymbol{y}$ to be physically meaningful, we assume it to be both orientation preserving ($\det(\nabla \boldsymbol{y}) > 0$, i.e., $\nabla \boldsymbol{y} \in \mathrm{GL}^+(d)$) and injective up to the boundary $\partial\Omega$.

Both in analytical discussions and in applications, it is often more convenient to consider the new relative position of a material point instead of its new absolute position. Thus, we introduce the so-called *displacement vector field*

$$\boldsymbol{u}\colon \overline{\Omega} \times [0,T] \to \mathbb{R}^d, \quad \boldsymbol{u}(x,t) \coloneqq \boldsymbol{y}(x,t) - x$$

with spatial gradient $\nabla \boldsymbol{u}(x,t) = \nabla \boldsymbol{y}(x,t) - \mathrm{I}$.

As we have already mentioned above, the deformation of the considered material is governed by external loads. These are composed of *volume* and *boundary* forces which we assume to be so-called *dead loads*, i.e., they do not depend on the body's deformation and can therefore be represented by their corresponding time-dependent force densities in the reference configuration Ω via

$$f_\Omega \colon \Omega \times [0,T] \to \mathbb{R}^d \quad \text{and} \quad f_{\Gamma_N} \colon \Gamma_N \times [0,T] \to \mathbb{R}^d .$$

These density functionals describe the force exerted per unit volume and per unit area, respectively, in the reference configuration. An illustration for the situation described above is given in Figure 2.1.

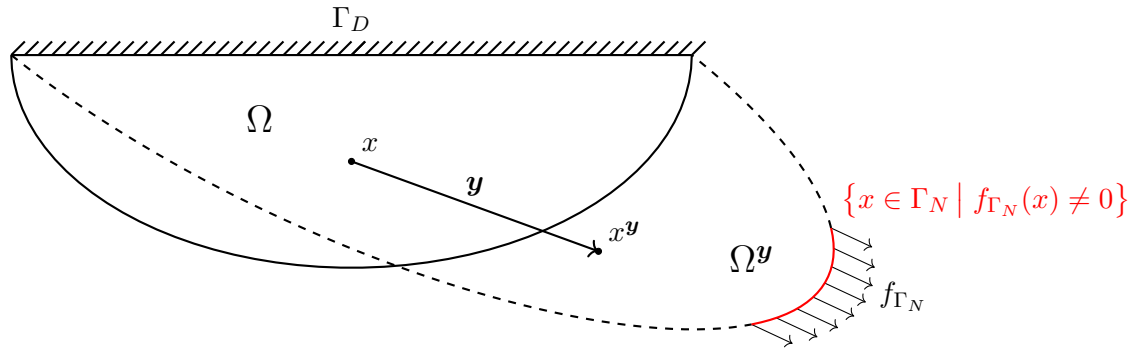


Figure 2.1: The deformation mapping $\boldsymbol{y} \colon \overline{\Omega} \to \mathbb{R}^2$ transforming the reference configuration Ω to the deformed configuration $\Omega^{\boldsymbol{y}}$ according to boundary conditions on Γ_D and Γ_N.

The movements initiated by either external or internal forces on the material can be classified into two types the nature of which will continue to play an important role over the course of what follows, cf. [6, Section 2.1.1]:

1. **Rigid body movements:** The distance between all points in the domain remains unchanged, i.e., the corresponding deformation field $\boldsymbol{y}$ satisfies

   $$\forall x, \tilde{x} \in \overline{\Omega},\, t \in [0,T] \colon \quad \big\| \boldsymbol{y}(x,t) - \boldsymbol{y}(\tilde{x},t) \big\| = \|x - \tilde{x}\| .$$

   As seen in [34, page 49], these can also be characterized as the sum of a translation and a rotation, i.e., continuous deformations of the form

   $$\boldsymbol{y}(x,t) = \boldsymbol{a}(t) + \mathrm{Q}(t)x \quad \text{with} \quad \boldsymbol{a}(t) \in \mathbb{R}^d,\, \mathrm{Q}(t) \in \mathrm{SO}(d,\mathbb{R})$$

   for all $x \in \overline{\Omega}$ and $t \in [0,T]$.

2. **Distorting movements:** The corresponding deformation field induces a change in distance of material points. This might cause parts of the body to be compressed while others might be stretched.

In order to now determine the response of the considered material body to distorting movements, we have to measure the so-called *strain* induced by these deformations. Distortion in general can be understood as the change in angles and lengths of material points over the

course of the movement. As described in [36, Section 2.1], to this end we from now on omit $t \in [0,T]$, consider a fixed point $x \in \Omega$ and two fibers of material particles emanating from x which we denote by Δx and δx. The corresponding fibers in the deformed configuration are given by

$$\Delta y \coloneqq \boldsymbol{y}(x+\Delta x) - \boldsymbol{y}(x) \quad \text{and} \quad \delta y \coloneqq \boldsymbol{y}(x+\delta x) - \boldsymbol{y}(x)\,.$$

Assuming sufficient differentiability of the deformation field, we consider the corresponding Taylor series in order to arrive at the expressions

$$\Delta y = \Delta x + (\nabla \boldsymbol{u})\Delta x + o\big(\|\Delta x\|\big) \quad \text{and} \quad \delta y = \delta x + (\nabla \boldsymbol{u})\delta x + o\big(\|\delta x\|\big)\,.$$

The difference of scalar products can thus be evaluated as

$$\begin{aligned}\langle \Delta y, \delta y\rangle_{\mathbb{R}^d} - \langle \Delta x, \delta x\rangle_{\mathbb{R}^d} = \langle (\nabla \boldsymbol{u})\Delta x, \delta x\rangle_{\mathbb{R}^d} + \langle (\nabla \boldsymbol{u})\delta x, \Delta x\rangle_{\mathbb{R}^d} \\ + \langle (\nabla \boldsymbol{u})\Delta x, (\nabla \boldsymbol{u})\delta x\rangle_{\mathbb{R}^d} + o\big(\|\delta x\|^2 + \|\Delta x\|^2\big)\end{aligned} \tag{2.2.1}$$

The geometrical interpretation of scalar products in $\mathbb{R}^3$ yields that, for rigid body movements, the above difference vanishes. For the case of a distortion, infitesimal expressions are of special interest. Thus, we will consider the limit of the above difference as the lengths of the fibers tend to zero. We set $h \coloneqq \max\{\|\delta x\|, \|\Delta x\|\}$ as well as $n \coloneqq \Delta x/h$ and $m \coloneqq \delta x/h$ as fixed vectors with directions that are independent of h. Dividing both sides of (2.2.1) by h^2 and taking the limit $h \to 0$ then gives

$$\lim_{h\to 0} \frac{\langle \Delta y, \delta y\rangle_{\mathbb{R}^d} - \langle \Delta x, \delta x\rangle_{\mathbb{R}^d}}{h^2} = 2\langle n, \mathbf{E}\, m\rangle_{\mathbb{R}^d}$$

where we identify $\mathbf{E}$ as the *strain tensor* associated with the displacement $\boldsymbol{u}$ defined by

$$\mathbf{E} \coloneqq \frac{1}{2}\big[\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T + (\nabla \boldsymbol{u})^T \nabla \boldsymbol{u}\big] = \frac{1}{2}\big[(\nabla \boldsymbol{y})^T \nabla \boldsymbol{y} - \mathrm{I}\big]\,. \tag{2.2.2}$$

Considering special cases and orientations for the up to now arbitrary fibers Δx and δx, we can deduce the interpretation of the diagonal components of $\mathbf{E}$ as half the net change in length (squared) of a material fiber originally oriented so that it points in the corresponding direction. Similarly, off-diagonal elements of the tensor give a measure of the change in angle between two fibers originally at right angles to each other and pointing towards the corresponding index directions. This interpretation characterizes diagonal elements as *direct strains* whereas off-diagonal elements are referred to as *shear strains*. Furthermore, $\mathbf{E} = 0$ implies that the body undergoes a rigid body motion.

For (infinitesimally) small displacements $\boldsymbol{u}$ and corresponding derivatives $\nabla \boldsymbol{u}$, we recognize that first order terms in $\mathbf{E}$ dominate those of second order. Thus, in the literature for so-called *small strain* scenarios, the strain within the body is replaced by a linearized version of the corresponding tensor at the vanishing displacement, leading to the *infinitesimal strain tensor*

$$\boldsymbol{\epsilon} \coloneqq \frac{1}{2}\big(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T\big)\,. \tag{2.2.3}$$

This expression is linear, symmetric and approximates $\mathbf{E}$ up to terms of $o\big(\|\nabla \boldsymbol{u}\|^2\big)$. These favorable properties make the small strain approach interesting both from an analytical and an application-oriented standpoint. Apparently, the approximation becomes poor for large displacements and is thus not generally suited for our interests.

With a measure for the distortion of the material in the form of the strain tensor $\mathbf{E}$ at hand, our next considerations are laid on the internal resistance counteracting the deformation and acting towards its resting position in the original shape. This internal resistance is often referred to as *stress*. In order to now obtain a mathematical expression for this material phenomenon, we consult the so called *stress principle of Euler and Cauchy* as stated in [14, Axiom 2.2-1]. It states the existence of a vector field

$$\boldsymbol{t}^{\boldsymbol{y}} \colon \overline{\Omega^{\boldsymbol{y}}} \times S_1(0) \to \mathbb{R}^d \tag{2.2.4}$$

such that both the *Axiom of Force Balance* and the *Axiom of Moment Balance*[5] are satisfied along every subdomain of the deformed configuration. The *stress vector* $\boldsymbol{t}^{\boldsymbol{y}}$ on its own then describes the force measured in units of pressure on an infinitesimal surface, to which the vector $n \in S_1(0)$ is normal at the point $x^{\boldsymbol{y}} \in \overline{\Omega^{\boldsymbol{y}}}$, as a reaction to the load on the body.

Of course, we want to express this crucial quantity of material response in terms of the forces acting on the body, even if for now only in the deformed configuration. Gladly, we find the following divergence-like dependence for a suitable tensor representing the stress vector, cf. [14, Theorem 2.3-1]:

Theorem 2.2.1: Cauchy's Theorem

Suppose that the internal and external force densities in the deformed configuration $\overline{\Omega^{\boldsymbol{y}}}$ are described by sufficiently regular vector fields $f^{\boldsymbol{y}}_{\Omega} \colon \Omega^{\boldsymbol{y}} \to \mathbb{R}^d$ and $f^{\boldsymbol{y}}_{\Gamma_N} \colon \Gamma^{\boldsymbol{y}}_N \to \mathbb{R}^d$. Then, there exists the so-called *Cauchy Stress Tensor*

$$\mathbf{T}^{\boldsymbol{y}} \colon \overline{\Omega^{\boldsymbol{y}}} \to \mathbb{R}^{d\times d}$$

such that the stress vector from (2.2.4) can be represented via

$$\forall\, (x^{\boldsymbol{y}}, n) \in \overline{\Omega^{\boldsymbol{y}}} \times S_1(0) \colon \quad \boldsymbol{t}^{\boldsymbol{y}}(x^{\boldsymbol{y}}, n) = \mathbf{T}^{\boldsymbol{y}}(x^{\boldsymbol{y}})n\,.$$

Furthermore, the stress tensor is related to the external forces in the form

$$\begin{aligned} \forall\, x^{\boldsymbol{y}} \in \Omega^{\boldsymbol{y}} \colon &\quad -\operatorname{div}^{\boldsymbol{y}}\big(\mathbf{T}^{\boldsymbol{y}}(x^{\boldsymbol{y}})\big) = f^{\boldsymbol{y}}_{\Omega}(x^{\boldsymbol{y}}) \ \text{ and} \\ \forall\, x^{\boldsymbol{y}} \in \Gamma^{\boldsymbol{y}}_N \colon &\quad \mathbf{T}^{\boldsymbol{y}}(x^{\boldsymbol{y}})\nu^{\boldsymbol{y}} = f^{\boldsymbol{y}}_{\Gamma_N}(x^{\boldsymbol{y}})\,, \end{aligned}$$

where $\operatorname{div}^{\boldsymbol{y}}$ denotes the divergence operator with respect to the deformed configuration and $\nu^{\boldsymbol{y}}$ is the unit outer normal vector in $x^{\boldsymbol{y}}$ on the deformed Neumann boundary $\Gamma^{\boldsymbol{y}}_N$.

The problem with the above divergence relation for the Cauchy Stress Tensor $\mathbf{T}^{\boldsymbol{y}}$ is that it is stated in the deformed configuration with $x^{\boldsymbol{y}} = \boldsymbol{y}(x)$ as a variable. As pointed out beforehand, we want to formulate the finite strain problem on the reference configuration and will thus transform $\mathbf{T}^{\boldsymbol{y}}$ to the so-called *first Piola-Kirchhoff Stress Tensor* $\mathbf{T} \colon \overline{\Omega} \to \mathbb{R}^{d\times d}$ via

$$\mathbf{T}(x) := \det\big(\nabla \boldsymbol{y}(x)\big)\mathbf{T}^{\boldsymbol{y}}(x^{\boldsymbol{y}})\nabla \boldsymbol{y}(x)^{-T} \quad \text{where} \quad x^{\boldsymbol{y}} = \boldsymbol{y}(x) \tag{2.2.5}$$

[5] Together, they state that any subdomain is in *static equilibrium*, in the sense that the *torsor* formed by elementary forces normal to its boundary and the body forces is equal to zero. In particular, the resultant vector (, i.e., the corresponding force,) and the resultant moment with respect to the origin vanish, cf. [14, page 61].

which in particular satisfies $\boldsymbol{t}(x,n) = \mathbf{T}(x)n$ for the *stress vector* $\boldsymbol{t}(x,n)$ formulated on the reference configuration. Generally, this tensor is non-symmetric which is often overcome by modifying it to the *second Piola-Kirchhoff Stress Tensor*, cf. [14, Section 2.5]. We, however, are interested in the differential equations formulated in terms of $\mathbf{T}$.

Formulation on the Reference Configuration

Similar to the transformation of the stress tensor, also internal and external forces as well as the divergence operator can be expressed via unknowns stemming from the reference configuration by using the determinant of the deformation gradient. Additonally, as we incorporate time-dependence to the problem by considering the body force $-\rho\ddot{\boldsymbol{u}}$ for the mass density function $\rho\colon \Omega \to [0,\infty[$, we thus arrive at the set of equations given by

$$\begin{aligned} \rho\ddot{\boldsymbol{u}} - \operatorname{div}(\mathbf{T}) &= f_\Omega \qquad \text{in} \quad \Omega \times [0,T]\,, \\ \mathbf{T}\nu &= f_{\Gamma_N} \qquad \text{on} \quad \Gamma_N \times [0,T] \end{aligned}$$

with now div the divergence operator with respect to $\overline{\Omega}$ and ν the unit outer normal vector in x on the reference Neumann boundary Γ_N.

Within the present treatise, we are interested in processes which happen very slowly in time such that it seems reasonable to disregard the $\rho\ddot{\boldsymbol{u}}$ term within the divergence equation above. This assumption is often referred to as the elasticity problem being *quasi-static*. As we then round out the formulation of the quasi-static *equilibrium equations for elasticity* by imposing Dirichlet boundary conditions on the remaining portion of $\partial\Omega$ given by Γ_D together with suitable initial values, we obtain the following boundary value problem:

$$\begin{aligned} -\operatorname{div}(\mathbf{T}) &= f_\Omega \qquad \text{in} \quad \Omega \times [0,T]\,, \\ \mathbf{T}\nu &= f_{\Gamma_N} \qquad \text{on} \quad \Gamma_N \times [0,T]\,, \\ \boldsymbol{u} &= 0 \qquad \text{on} \quad \Gamma_D \times [0,T]\,. \end{aligned} \tag{2.2.6}$$

In its definition in (2.2.5), we have omitted the explicit dependence of $\mathbf{T}$ on the deformation gradient $\nabla\boldsymbol{y}$. Commonly in the literature, cf. e.g. [14, Chapter 3], a material is referred to as *elastic* if and only if each Piola-Kirchhoff stress tensor can be expressed in terms of the material point $x \in \overline{\Omega}$ and the corresponding deformation gradient $\nabla\boldsymbol{y}(x)$ through a so-called *constitutive equation* of the form

$$\forall\, x \in \overline{\Omega}\colon \quad \mathbf{T}(x) = \widehat{\mathbf{T}}\big(x, \nabla\boldsymbol{y}(x)\big)$$

where the *response function* $\widehat{\mathbf{T}}\colon \overline{\Omega} \times \mathrm{GL}^+(d) \to \mathbb{R}^{d\times d}$ characterizes the elastic material.

We are particularly interested in this representation of the first Piola-Kirchhoff tensor since we want to consider so-called *hyperelastic materials* as introduced in [14, Chapter 4]. The latter are characterized by the existence of a continuously differentiable *stored energy function* $\hat{W}\colon \overline{\Omega} \times \mathrm{GL}^+(d) \to \mathbb{R}$ such that the relation

$$\forall\, x \in \overline{\Omega}\,,\ \mathrm{F} \in \mathrm{GL}^+(d)\colon \quad \widehat{\mathbf{T}}(x,\mathrm{F}) = \mathrm{D}_\mathrm{F}\hat{W}(x,\mathrm{F}) \tag{2.2.7}$$

holds. If this is the case, and if the applied forces are *conservative*[6], solving the boundary value problem from (2.2.6) is formally equivalent to finding a stationary point of a *total energy*

[6]Conservative body forces can be expressed as the Gâteaux derivative of an integrated functional. The corresponding integrand is then called the potential of the applied body force.

functional $\mathcal{E}$. For *dead loads*[7], we even have the representation

$$\mathcal{E}(\psi, \nabla\psi) = \int_\Omega \hat{W}\big(x, \nabla\psi(x)\big)\,\mathrm{d}x - \Big(\int_\Omega \langle f_\Omega(x), \psi(x)\rangle_{\mathbb{R}^d}\,\mathrm{d}x + \int_{\Gamma_N} \langle f_{\Gamma_N}, \psi\rangle_{\mathbb{R}^d}\,\mathrm{d}S\Big) \tag{2.2.8}$$

for all *admissible deformations* $\psi\colon \overline{\Omega} \to \mathbb{R}^d$ with $\det(\nabla\psi) > 0$ and $\psi(x) = x$ for all $x \in \Gamma_D$. Time-dependence of all occurring functionals can be inserted without problems in our quasi-static case.

As formulated in [14, Theorem 4.1-1], this finally provides us with a variational principle leading to the differential identity

$$\forall\, t \in [0,T]\colon\ \mathrm{D}_\psi \mathcal{E}\big(t, \boldsymbol{y}(t), \nabla\boldsymbol{y}(t)\big)\theta = 0 \tag{2.2.9}$$

for all sufficiently smooth maps $\theta\colon \overline{\Omega} \to \mathbb{R}^d$ vanishing on Γ_D which is formally equivalent to the boundary value problem from (2.2.6). The identity from (2.2.9) will later on also be a central part of the rate-independent formulation of the elasto-plastic problem.

2.2.2 Yield Surfaces, Flow Rules, and Rate-Independence

While the theory introduced in Section 2.2.1 above adequately describes the elastic behavior of the materials which we want to consider within our simulations later on, we will go beyond the theory of elasticity for an adequate description of the response to external loads. To this end, we will now give a comprehensive overview of both the motivation and theory behind plastic behavior. The following physical background and motivation is based on the deliberations from [36, Chapter 3].

In order to depict the fundamental features of elasto-plastic materials, it is reasonable to consider the simple situation of uni-axial stress in a body. From an application-oriented standpoint, one might consider a thin rod to which a force f_{Γ_N} is applied at each end, acting in different directions. The rather simple setup of this experiment is illustrated in Figure 2.2.

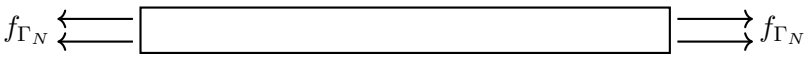


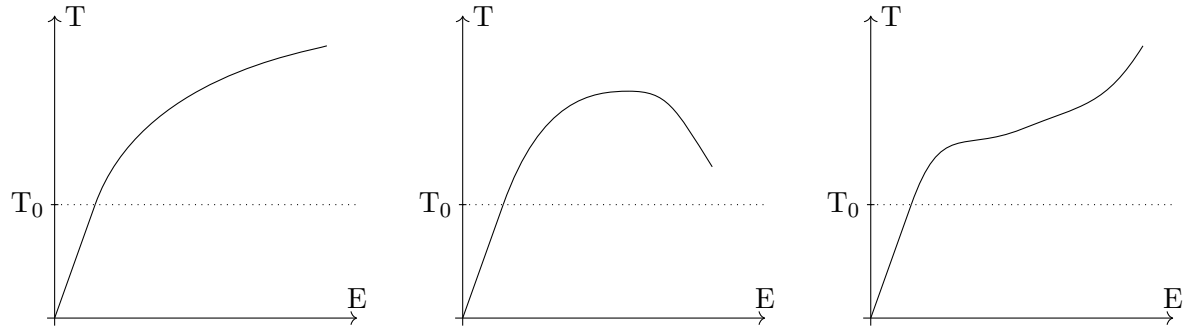
Figure 2.2: An elasto-plastic rod in uni-axial stress T caused by the boundary force f_{Γ_N}.

What we are interested in now is the response of the material to the uni-axial stress applied to each end of the rod, i.e., the so-called *stress-strain relationship*. For this one-dimensional scenario, we denote the stress ensuing from occurring external and internal forces by T, cf. (2.2.5). In the above example, an adequate experiment is to gradually increase the force acting on the rod leading to a change of length in the rod and thereby a corresponding increase in strain E, cf. (2.2.2). This dependence allows to record the history of behavior during a program of loading and is suited for the explanation of plastic contrary to purely elastic material response.

Non-Linearity and Path-Dependence

Typical graphs of this stress-strain dependence are given in Figure 2.3. Up to the so-called *initial yield stress* T_0, we have linear elastic behavior in all three considered cases. If the

[7] As mentioned beforehand, the force density of dead loads is independent of the particular deformation $\boldsymbol{y}$.



(a) Hardening behavior of plastic materials (, e.g. rubber).

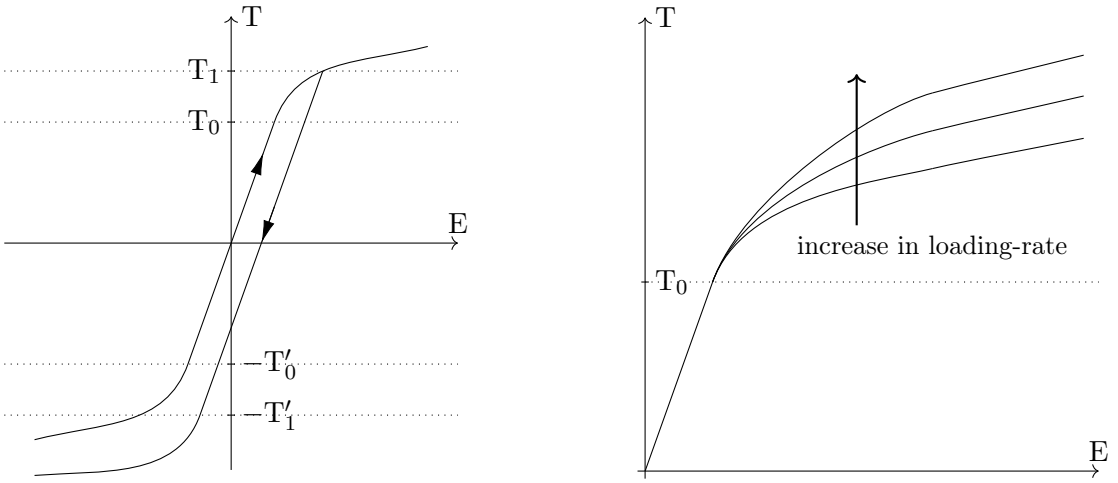
(b) Softening behavior of plastic materials (, e.g. concrete).

(c) Stiffening behavior of plastic materials (, e.g. metals).

Figure 2.3: Relationship of the uni-axial stress T induced by an increasing external force and the corresponding strain E within an elasto-plastic rod for different materials.

applied force, and hence also the stress, is increased further, the behavior deviates from the linear relation which is often used for an approximation in small strain elastic theory, cf. [14, Section 6.3] or [36, 49, 101]. A widely spread feature is the decrease in the slope of the corresponding curve which will continue until eventually a variety of phenomena may take place. Depending on the application in question and on the range of stress which is expected to be experienced, *hardening* (see Figure 2.3a, e.g. rubber,), *softening* (see Figure 2.3b, e.g. in soil or concrete,), and *stiffening* (see Figure 2.3c, e.g. in some metals,) are common.

Up to now, our deliberations concerning plastic behavior only illustrate the inherent *non-linearity* of the stress-strain relationship. This is by itself not a distinctive property of plasticity but might also be modeled by employing non-linear purely elastic models. The feature which rules out the latter possibility is the one of *irreversibility* or *path-dependence* of plastic deformations. This phenomenon describes the circumstance that – unlike in the case of elasticity – the state of strain does not revert to its original state upon removal of applied forces.



(a) Irreversibility (or path-dependence) and elastic unloading.

(b) Loading at increasing rates for a rate-dependent material.

Figure 2.4: Irreversibility and possible rate-dependence during loading for plastic materials.

This characteristic feature of plastic material response can be retraced in Figure 2.4a. Here, we also allow for negative values of T, i.e., compression of the material exhibiting similar features as described above. Reversing the direction of loading at some stress $\mathrm{T}_1 > \mathrm{T}_0$ does not cause the state to follow back the original curve to its initial configuration. Instead the material behaves elastically and evolves along the straight line with the same slope as in the original loading process. This phenomenon is referred to as *elastic unloading*.

Additionally, plastic deformation beyond the initial yield stress T_0 (or $-\mathrm{T}_0'$) also affects the so-called *elastic region* of the material. While the *initial elastic range* $]-\mathrm{T}_0', \mathrm{T}_0[$ includes the original state, this might not be the case for subequent elastic ranges like $]-\mathrm{T}_1', \mathrm{T}_1[$ which are only accessible by plastic deformation previously haven taken place.

While still on the topic of stress-strain relationships, let us shortly remark on the role and physical background of *rate-independence* within elasto-plastic behavior. Therefore consider the experiment from above to be repeated multiple times and each time the rate, at which the external force is applied, is different. In many applications, it is observed that – while the elastic response remains unchanged – the plastic behavior of the material differs with the rate of loading. This phenomenon of rate-dependence as illustrated in Figure 2.4b will be neglected within the present treatise in order to take advantage of the rich theory from Section 2.1.

Introduction of the Plastic Variable

In order to now get a grasp on the plastic phenomena of non-linearity and path-dependence described above, it seems natural to introduce a new *plastic variable* $\mathbf{P}$ to our description of material response. For the definition of this variable which measures the state of plastic deformation, it is intuitive to decompose the total strain into a purely elastic and a purely plastic component. This approach prevails in particular in small strain theory, cf. [36, 45, 59, 75], where an additive decomposition of the infinitesimal strain tensor from (2.2.3) via

$$\boldsymbol{\epsilon} = \boldsymbol{e} + \boldsymbol{p}$$

is proposed. Thermodynamic considerations support this approach [36, Section 3.2] and it is adequate for the vector space setting in which small strain theory is often formulated.

For the consideration of large deformations and thereby finite strain elasto-plasticity, a manifold-based description within a Lie-group setting turns out to be advantageous, cf. [63]. In order to enable this formulation of the underlying theory, we use a *multiplicative decomposition* of the deformation gradient $\nabla \boldsymbol{y} \colon \overline{\Omega} \times [0,T] \to \mathrm{GL}^+(d)$ into an *elastic part* $\mathbf{F}_{el} \colon \overline{\Omega} \times [0,T] \to \mathrm{GL}^+(d)$ and a *plastic part* $\mathbf{F}_p \colon \overline{\Omega} \times [0,T] \to \mathrm{SL}(d)$ via

$$\nabla \boldsymbol{y} =: \mathbf{F} = \mathbf{F}_{el}\mathbf{F}_p \,. \tag{2.2.10}$$

In the following, we will again focus on a conceptional deduction of equations describing the plastic behavior motivated beforehand. Thus, we will often notationally identify the mappings introduced above with their corresponding images for fixed $(x,t) \in \overline{\Omega} \times [0,T]$ and also omit explicit time-dependence of occurring functionals.

The multiplicative split from (2.2.10) for the manifold-based description can be illustrated by the existence of the so-called *intermediate configuration* which only depicts the internal plastic state of the material. The deformation gradient as a whole can then be understood as a concatenation of separate plastic and elastic deformation mappings, cf. [23, 47, 54] for first

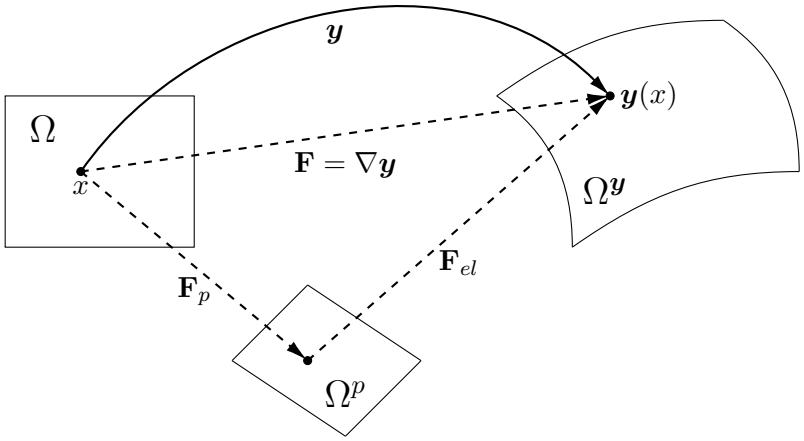


Figure 2.5: Illustration of the multiplicative split of the deformation gradient and introduction of the intermediate configuration Ω^p.

approaches to such formulations or [10, 91] for a brief historical overview. This understanding is visualized in Figure 2.5.

The plastic variable itself can from here on be interpreted differently. While for analytical arguments as in [63], it is often advantageous to consider $\mathbf{P} \coloneqq \mathbf{F}_p^{-1}$,[8] application-based approaches prefer to use $\mathbf{P} \coloneqq \mathbf{F}_p$ instead, cf. [66, Section 4.2.1] or our application problem derived in Section 2.2.4. The resulting descriptions in the Lie-group scenario are equivalent since in a manifold setting choosing between these alternatives for $\mathbf{P}$ can be viewed analogous to considering different signs $\pm$ in vector spaces. For now, we will go with the former of these choices in order to deduce the rate-independent formulation of finite strain plasticity.

The first step towards the generalization of the theory from Section 2.2.1 to plastic behavior is the natural extension of the stored energy functional from (2.2.7) to incorporating a split dependence on the elastic deformation gradient $\mathbf{F}_{el}$ and the plastic variable $\mathbf{P}$ at some $(x,t) \in \overline{\Omega} \times [0,T]$ via

$$\hat{W} \colon \overline{\Omega} \times \mathrm{GL}^+(d) \times \mathrm{SL}(d) \to \mathbb{R} \,, \quad (x, \mathrm{F}, \mathrm{P}) \mapsto \hat{W}(x, \mathrm{F}, \mathrm{P})$$

where we will mostly omit the explicit dependence on $x \in \overline{\Omega}$ for the sake of notational simplicity.

Yield Functions, Surfaces, and the Plastic Flow Rule

The characteristic plastic phenomena described beforehand and the evolution of the plastic variable $\mathbf{P}$ can then be modeled by using so-called *yield functions* and *surfaces*. The yield function $Y \colon \mathrm{SL}(d) \times \mathbb{R}^{d\times d} \to \mathbb{R}$, $(\mathrm{P}, \mathrm{Q}) \mapsto Y(\mathrm{P}, \mathrm{Q})$, helps us to characterize the elastic region, i.e., the interval $]-\mathrm{T}_0', \mathrm{T}_0[$ in the above example, via the sublevel set $\{Y < 0\}$, and its boundary, the yield surface where plastic deformation takes place, via the level set $\{Y = 0\}$. Thus, the yield function determines the plastic behavior of the considered material and is assumed to depend on $\mathbf{P}$ as well as on the so-called *plastic back-stress*

$$\mathbf{Q} \colon \overline{\Omega} \times [0,T] \to \mathbb{R}^{d\times d} \,, \quad \mathbf{Q}(x,t) = -\mathrm{D}_{\mathrm{P}} \hat{W}\big(x, \mathbf{F}_{el}(x,t), \mathbf{P}(x,t)\big) \,.$$

[8] In this case, we do not really consider the inverse of the mapping $\mathbf{F}_p \colon \overline{\Omega} \times [0,T] \to \mathrm{SL}(d)$ but interpret $\mathbf{P} \colon \overline{\Omega} \times [0,T] \to \mathrm{SL}(d)$, $\mathbf{P}(x,t) \coloneqq \mathbf{F}_p(x,t)^{-1}$, as the corresponding pointwise inverse matrix.

The latter can be seen as the conjugate variable to the Piola-Kirchhoff stress $\mathbf{T}$ characterized by (2.2.7). For the sake of an easier analysis as in [66], we will now assume that the yield function only depends on $\mathbf{P}$ and $\mathbf{Q}$ through the tensor $\bar{\mathbf{Q}} \coloneqq \mathbf{P}^T\mathbf{Q}$, i.e., we have that $Y(\mathbf{P},\mathbf{Q}) = \hat{Y}(\bar{\mathbf{Q}})$ holds for an adequate function $\hat{Y}\colon \mathbb{R}^{d\times d} \to \mathbb{R}$, $\bar{Q} \mapsto \hat{Y}(\bar{Q})$. With this formulation at hand, the *principle of maximal dissipation* provides us with the so-called *associated flow rule* for the evolution of $\mathbf{P}$, cf. [99, 100, 119]. As discussed previously in Section 2.1.2, the principle often formulated via (2.1.13) constitutes a fundamental assertion within the physical description and has crucial implications for the setting we find ourselves in. In particular, the flow rule can then be formulated in *Karush-Kuhn-Tucker form* via

$$\mathbf{P}^{-1}\dot{\mathbf{P}} = \lambda \mathrm{D}_{\bar{Q}}\hat{Y}(\bar{\mathbf{Q}})\,, \quad \text{with } \lambda \geq 0\,,\ \hat{Y}(\bar{\mathbf{Q}}) \leq 0 \text{ and } \lambda\hat{Y}(\bar{\mathbf{Q}}) = 0\,. \tag{2.2.11}$$

We, however will reformulate this statement for fixed $(x,t) \in \overline{\Omega}\times[0,T]$ and to this end introduce the set of *admissible generalized stresses* $\mathbb{Q}$ together with its characteristic function $\mathcal{X}_{\mathbb{Q}}\colon \mathbb{R}^{d\times d} \to [0,\infty]$ by

$$\mathbb{Q} \coloneqq \left\{\bar{\mathrm{Q}} \in \mathbb{R}^{d\times d} \,\middle|\, \hat{Y}(\bar{\mathrm{Q}}) \leq 0\right\} \quad \text{and} \quad \mathcal{X}_{\mathbb{Q}}(\bar{\mathrm{Q}}) \coloneqq \begin{cases} 0 & \text{, if } \bar{\mathrm{Q}} \in \mathbb{Q} \\ \infty & \text{, otherwise.} \end{cases}$$

Here, we assume convexity of $\mathbb{Q}$ together with the existence of $r_1, r_2 > 0$ such that $B_{r_1}(0) \subset \mathbb{Q} \subset B_{r_2}(0)$. The set $\mathbb{Q}$ mirrors the abstract elasticity domain from (2.1.11). In particular, also $\mathcal{X}_{\mathbb{Q}}$ is thereby convex and we can reformulate (2.2.11) via

$$\mathbf{P}^{-1}\dot{\mathbf{P}} \in \partial\mathcal{X}_{\mathbb{Q}}(\bar{\mathbf{Q}}) = N_{\bar{\mathbf{Q}}}\mathbb{Q} \tag{2.2.12}$$

where the latter identity involves the *outer normal cone* which in general is defined via

$$N_z C \coloneqq \left\{z^* \in Z^* \,\middle|\, \forall\, \hat{z} \in C\colon \langle z^*, z - \hat{z}\rangle \leq 0\right\}$$

for a closed convex set C in the point $z \in C$ within a Banach space Z with dual space Z^*. Now, we define the *dissipation functional* as the convex conjugate of $\mathcal{X}_{\mathbb{Q}}$ by

$$\Delta\colon \mathbb{R}^{d\times d} \to [0,\infty]\,, \quad \Delta(\xi) \coloneqq (\mathcal{X}_{\mathbb{Q}})^*(\xi) = \sup\left\{\langle \xi, \eta\rangle \,\middle|\, \eta \in \mathbb{Q}\right\}$$

and conclude convexity as well as homogeneity of degree 1 of said mapping. Classical duality theory for convex functions (, cf. [90, Proposition 11.3],) thus provides us with the equivalence of $\xi \in \partial\mathcal{X}_{\mathbb{Q}}(\bar{\mathrm{Q}})$ and $\bar{\mathrm{Q}} \in \partial\Delta(\xi)$. Using $\xi = \mathbf{P}^{-1}\dot{\mathbf{P}}$, we can thereby reformulate the flow rule from (2.2.12) in terms of

$$\hat{\Delta}\colon \big(\mathrm{SL}(d)\big)^2 \to [0,\infty]\,, \quad \hat{\Delta}(\mathrm{A},\mathrm{B}) \coloneqq \Delta(\mathrm{A}^{-1}\mathrm{B})$$

via $\mathbf{Q} \in \partial^{\mathrm{B}}\hat{\Delta}(\mathbf{P},\dot{\mathbf{P}})$. The definition of $\mathbf{Q}$ then leads to the final pointwise version of the flow rule as the following differential inclusion:

$$0 \in \mathrm{D}_{\mathrm{P}}\hat{W}(x,\mathbf{F}_{el},\mathbf{P}) + \partial^{\mathrm{B}}\hat{\Delta}(\mathbf{P},\dot{\mathbf{P}}) \subset \mathbb{R}^{d\times d}\,.$$

In order to now obtain the corresponding field equation, we consider the integrated total energy functional from (2.2.8) (which now accordingly also depends on the plastic variable $\mathbf{P}$) together with the *integrated dissipation functional*

$$\hat{\Delta}_\Omega\colon \big(\Omega \to \mathrm{SL}(d)\big)^2 \to [0,\infty]\,, \quad \hat{\Delta}_\Omega(\mathbf{A},\mathbf{B}) \coloneqq \int_\Omega \hat{\Delta}\big(\mathbf{A}(x),\mathbf{B}(x)\big)\,\mathrm{d}x$$

and thus obtain the *plastic flow rule*:

$$\forall\, t \in [0,T]\colon\ 0 \in \mathrm{D}_{\mathrm{P}}\mathcal{E}(t, \boldsymbol{y}, \mathbf{P}) + \partial^{\mathbf{B}}\hat{\Delta}_{\Omega}(\mathbf{P}, \dot{\mathbf{P}})\,. \tag{2.2.13}$$

Together with the variational formulation of the elastic equilibrium equation from (2.2.9), this constitutes a complete description of finite strain elasto-plasticity. Combining the differential inclusions (2.2.9) and (2.2.13) into one, we achieve

$$\forall\, t \in [0,T]\colon\ 0 \in \mathrm{D}_{\boldsymbol{q}}\mathcal{E}\big(t, \boldsymbol{q}(t)\big) + \partial^{v}\mathcal{R}\big(\boldsymbol{q}(t), \dot{\boldsymbol{q}}(t)\big)$$

for the state variable $\boldsymbol{q} \coloneqq (\boldsymbol{y}, \mathbf{P})$ and the dissipation potential $\mathcal{R} \coloneqq \hat{\Delta}_{\Omega}$ which is exactly the differential inclusion from (2.1.5) for rate independent systems. The above formulation with separated (2.2.9) and (2.2.13) can also be thought of as the coupled system from (2.1.17) which allows for the interpretation of the deformation field $\boldsymbol{y}$ as the non-dissipative counterpart to the dissipative plastic strain variable $\mathbf{P}$. This concludes the rate-independent formulation of elasto-plasticity and enables the discussion of the existence of corresponding energetic solutions in what follows.

2.2.3 Existence of Energetic Solutions

Now that we have established the transition from continuum mechanics to a formulation of finite strain plasticity within a rate independent system, the next goal is to lay out a general framework of assumptions on the stored energy functional and the dissipation distance which allows for energetic solutions as characterized in Section 2.1.2. We will also elaborate on how the assumptions formulated here transfer to their abstract counterparts from the general existence theory and thus give an idea how the proof of existence is structured without going too much into technical detail. For a detailed elaboration, we refer to [66, Section 4.2.1].

As we have already mentioned beforehand, we will also here employ the multiplicative decomposition of the deformation gradient but this time around interpret the plastic variable as $\mathbf{P} \coloneqq \mathbf{F}_p$. While a general assumption like $\mathbf{P}(x,t) \in \mathrm{GL}^+(d)$ is possible, we will continue to use $\mathbf{P}(x,t) \in \mathrm{SL}(d)$ together with $\mathbf{F}_{el}(x,t) \in \mathrm{GL}^+(d)$. In addition to $\mathbf{P}$, we include further plastic variables like *hardening variables* and *slip strains* that are combined into a vector-valued mapping $\boldsymbol{\Pi}\colon \overline{\Omega} \times [0,T] \to \mathbb{R}^m$ for $m \in \mathbb{N}$.

As for the dissipative components before, we now write $z = (\mathrm{P}, \Pi) \in \mathrm{SL}(d) \times \mathbb{R}^m$ in general, $\boldsymbol{z} = (\mathbf{P}, \boldsymbol{\Pi})$ for the corresponding mappings on $\overline{\Omega} \times [0,T]$, and use A as a placeholder for images of the (spacial) gradient contributions

$$\nabla \boldsymbol{z}\colon \overline{\Omega} \times [0,T] \to \mathbb{R}^{d\times d\times d} \times \mathbb{R}^{m\times d} =: \mathrm{L}^{(d,m)}\,, \quad \nabla \boldsymbol{z}(x,t) \coloneqq \big(\nabla \mathbf{P}(x,t), \nabla \boldsymbol{\Pi}(x,t)\big)\,.$$

The latter also characterize the models of plasticity which we are considering here, often referred to as *gradient theories* or *gradient plasticity*.

With the main unknowns in place, we consider the *stored energy density* W as the sum of an elastic part W_{el} and a part W_{hd} including hardening and regularizing terms, i.e., we have

$$W(x, \mathrm{F}, \mathrm{P}, \Pi, \mathrm{A}) = W_{\mathrm{el}}(x, \mathrm{F}, \mathrm{P}) + W_{\mathrm{hd}}(x, \mathrm{P}, \Pi, \mathrm{A}) \tag{2.2.14}$$

for all $x \in \Omega, \mathrm{F} \in \mathrm{GL}^+(d), \mathrm{P} \in \mathrm{SL}(d), \Pi \in \mathbb{R}^m$ and $\mathrm{A} \in \mathrm{L}^{(d,m)}$.

For notational simplicity, we will assume time-dependent Dirichlet-data $\boldsymbol{y}_D\colon [0,T] \times \Gamma_D \to \mathbb{R}^3$ to drive the process instead of volume and surface forces. The changes to be made in order

to include non-trivial forces can be retraced in [66, Remark 4.2.5]. Concisely, volume and surface forces have to be chosen such that for

$$\langle \boldsymbol{\ell}(t), \tilde{\boldsymbol{y}} \rangle \coloneqq \int_\Omega \langle f_\Omega(x,t), \tilde{\boldsymbol{y}}(x) \rangle_{\mathbb{R}^d} \,\mathrm{d}x + \int_{\Gamma_N} \langle f_{\Gamma_N}(t), \tilde{\boldsymbol{y}} \rangle_{\mathbb{R}^d} \,\mathrm{d}S \tag{2.2.15}$$

the functional inclusion $\boldsymbol{\ell} \in W^{1,1}\big(0,T; W^{1,p_{\mathrm{df}}}(\Omega;\mathbb{R}^d)^*\big)$ holds with $p_{\mathrm{df}} > d$ to be determined later. For the function governing time-dependent Dirichlet-data, we assume expandability from the Dirichlet boundary of non-zero measure to $\mathbb{R}^3$ such that the derivative expressions $\nabla \boldsymbol{y}_D, \nabla \dot{\boldsymbol{y}}_D, (\nabla \boldsymbol{y}_D)^{-1}$ are bounded and continuous on $[0,T] \times \mathbb{R}^d$. Thus, we can write the desired deformation field as the composition

$$\boldsymbol{y}(x,t) = \boldsymbol{y}_D\big(\tilde{\boldsymbol{y}}(x,t), t\big) \quad \text{with } \tilde{\boldsymbol{y}}(\cdot,t) \in \mathcal{Y}$$

for the *space of admissible deformations* $\mathcal{Y}$ given as

$$\mathcal{Y} \coloneqq \big\{ \tilde{\boldsymbol{y}} \in \boldsymbol{Y} \,\big|\, \forall\, x \in \Gamma_D \colon \tilde{\boldsymbol{y}}(x) = x \big\} \quad \text{with } \boldsymbol{Y} \coloneqq W^{1,p_{\mathrm{df}}}\big(\Omega;\mathbb{R}^d\big) \tag{2.2.16}$$

where we will specify $d < p_{\mathrm{df}} < \infty$ later. With this definition at hand, we characterize the *domain of the internal variable* by

$$\begin{aligned} &\mathcal{Z} \coloneqq \big\{ (\mathbf{P}, \boldsymbol{\Pi}) \in \boldsymbol{Z} \,\big|\, \mathbf{P} \in \mathrm{SL}(d) \text{ a.e. in } \Omega \big\} \\ &\text{with } \boldsymbol{Z} \coloneqq \big[L^{p_{\mathrm{pl}}} \cap W^{1,p_{\mathrm{gr}}}\big](\Omega;\mathbb{R}^{d\times d}\big) \times \big[L^{p_{\mathrm{hd}}} \cap W^{1,p_{\mathrm{gr}}}\big](\Omega;\mathbb{R}^m\big) \end{aligned} \tag{2.2.17}$$

again with $p_{\mathrm{pl}}, p_{\mathrm{gr}}, p_{\mathrm{hd}} \in]1,\infty[$ to be determined later. Omitting the tilde on the deformation field above, we recognize the *stored energy functional* $\mathcal{E}$ and the *dissipation distance* $\mathcal{D}$ as

$$\begin{aligned} \mathcal{E}(t,\boldsymbol{y},\boldsymbol{z}) &\coloneqq \int_\Omega W\big(t,x,\nabla \boldsymbol{y}_D\big(t,\boldsymbol{y}(x,t)\big)\nabla \boldsymbol{y}(x,t)\mathbf{P}(x,t)^{-1}, \boldsymbol{z}(x,t), \nabla \boldsymbol{z}(x,t)\big)\,\mathrm{d}x\,, \\ \mathcal{D}(\boldsymbol{z}_1,\boldsymbol{z}_2) &\coloneqq \int_\Omega D\big(x,\boldsymbol{z}_1(x,t),\boldsymbol{z}_2(x,t)\big)\,\mathrm{d}x\,. \end{aligned} \tag{2.2.18}$$

Going on with formulating assumptions on the underlying mappings and domains, we demand D to be an extended quasi-distance with

$$\begin{aligned} &D\colon \Omega \times \big(\mathrm{SL}(d)\times\mathbb{R}^m\big)^2 \to [0,\infty[\text{ is a normal integrand,} \\ &\forall\, x \in \Omega,\ z_1, z_2 \in \mathrm{SL}(d)\times\mathbb{R}^m\colon \quad D(x,z_1,z_2) = 0 \iff z_1 = z_2\,, \\ &\forall\, x \in \Omega,\ z_1, z_2, z_3 \in \mathrm{SL}(d)\times\mathbb{R}^m\colon \quad D(x,z_1,z_3) \le D(x,z_1,z_2) + D(x,z_2,z_3) \end{aligned} \tag{2.2.19}$$

where a normal integrand $a\colon \Omega\times\mathbb{R}^m \to \mathbb{R}_\infty$ is characterized by lower semi-continuity of $a(x,\cdot)$ for almost all $x \in \Omega$ and measurability of a as a whole.

As far as the stored energy density $W\colon \Omega \times \mathrm{GL}^+(d) \times \big(\mathrm{SL}(d)\times\mathbb{R}^m\big) \times \mathrm{L}^{(d,m)} \to]-\infty,\infty]$ is concerned, we have to put in some more deliberations. Firstly, we demand W to be *polyconvex* together with a particular lower bound for *coercivity* on its domain. To this end, we introduce the function $\mathbb{M}\colon \mathbb{R}^{d\times d} \to \mathbb{R}^{\mu_d}$ with $\mu_d \coloneqq \binom{2d}{d} - 1$ which maps a matrix to all its *minors* (subdeterminants). Then, we can formulate the conditions

$$\begin{aligned}
&\exists\, \mathbb{W}\colon \Omega \times \mathbb{R}^{\mu_d} \times \mathrm{SL}(d) \times \mathbb{R}^m \times \mathrm{L}^{(d,m)} \to]-\infty,\infty]\colon \\
&\qquad \text{(i)} \quad \mathbb{W} \text{ is a normal integrand,} \\
&\qquad \text{(ii)} \quad \forall\,(x,\mathrm{F},z,\mathrm{A})\colon W(x,\mathrm{F},z,\mathrm{A}) = \mathbb{W}\big(x,\mathbb{M}(\mathrm{F}),z,\mathrm{A}\big) \\
&\qquad \text{(iii)} \quad \forall\,(x,z)\colon \mathbb{W}\big(x,\cdot,z,\cdot\big)\colon \mathbb{R}^{\mu_d} \times \mathrm{L}^{(d,m)} \to]-\infty,\infty] \text{ is convex,}
\end{aligned} \tag{2.2.20a}$$

$$\begin{aligned}
&\exists\, c>0,\ h \in L^1(\Omega),\ p_{\mathrm{el}},\ p_{\mathrm{pl}},\ p_{\mathrm{hd}},\ p_{\mathrm{gr}} > 1\ \forall\,(x,\mathrm{F},\mathrm{P},\Pi,\mathrm{A}) \in \mathrm{dom}W\colon \\
&\qquad W(x,\mathrm{F},\mathrm{P},\Pi,\mathrm{A}) \geq h(x) + c\big(\|\mathrm{F}\|^{p_{\mathrm{el}}} + \|\mathrm{P}\|^{p_{\mathrm{pl}}} + \|\Pi\|^{p_{\mathrm{hd}}} + \|\mathrm{A}\|^{p_{\mathrm{gr}}}\big)\,.
\end{aligned} \tag{2.2.20b}$$

The polyconvexity assumption (2.2.20a), introduced in [5], together with the boundedness and continuity of $\nabla \boldsymbol{y}_D$, $\nabla \dot{\boldsymbol{y}}_D$ and $(\nabla \boldsymbol{y}_D)^{-1}$ provides weak lower semi-continuity of the stored energy functional $\mathcal{E}(t,\cdot)\colon \boldsymbol{Y} \times \boldsymbol{Z} \to \mathbb{R}$ as defined in (2.2.18) at every time $t \in [0,T]$ if the relations

$$\frac{1}{p_{\mathrm{el}}} + \frac{1}{p_{\mathrm{pl}}} = \frac{1}{p_{\mathrm{df}}} < \frac{1}{d}\,, \quad p_{\mathrm{hd}} > 1\,, \quad \text{and} \quad p_{\mathrm{gr}} > 1 \tag{2.2.21}$$

hold for the Lebesgue exponents. This implication together with the corresponding proof can be retraced in [66, Proposition 4.1.4].

While the assumptions in (2.2.20) only concern the stored energy density itself, we also have to demand certain bounds and continuity estimates on the derivatives of W with respect to elastic deformation placeholders F close to the identity. For this reason, we define a *modulus of continuity* ω as a non-decreasing function with $\omega(\rho) \to 0$ for $\rho \xrightarrow{>} 0$. Now, we are in the position to formulate

$$\begin{aligned}
&\exists\, c_0^W \in \mathbb{R},\ c_1^W > 0,\ \delta > 0, \text{ modulus of continuity } \omega\colon]0,\delta[\to]0,\infty[\\
&\quad \forall\,(x,\mathrm{F},z,\mathrm{A}) \in \mathrm{dom}W\ \forall\,\mathrm{N} \in \mathcal{N}_\delta \coloneqq \big\{\mathrm{N} \in \mathbb{R}^{d\times d} \,\big|\, \|\mathrm{N}-\mathrm{I}\| < \delta\big\}\colon \\
&\qquad W(x,\cdot,z,\mathrm{A}) \text{ is differentiable on } \mathcal{N}_\delta\mathrm{F} \text{ and} && \text{(2.2.22a)} \\
&\qquad \big|\mathrm{D}_{\mathrm{F}} W(x,\mathrm{F},z,\mathrm{A})\mathrm{F}^T\big| \leq c_1^W\big(W(x,\mathrm{F},z,\mathrm{A}) + c_0^W\big) && \text{(2.2.22b)} \\
&\qquad \big|\mathrm{D}_{\mathrm{F}} W(x,\mathrm{F},z,\mathrm{A})\mathrm{F}^T - \mathrm{D}_{\mathrm{F}} W(x,\mathrm{NF},z,\mathrm{A})(\mathrm{NF})^T\big| \\
&\qquad\qquad \leq \omega\big(\|\mathrm{N}-\mathrm{I}\|\big)\big(W(x,\mathrm{F},z,\mathrm{A}) + c_0^W\big) && \text{(2.2.22c)}
\end{aligned}$$

where we denoted images of F by $\mathcal{N}_\delta\mathrm{F} \coloneqq \{\mathrm{NF} \,|\, N \in \mathcal{N}_\delta\}$. To give a short interpretation, we can see (2.2.22b) as a *multiplicative stress control* since the so-called *Kirchhoff stress tensor* $\mathrm{D}_{\mathrm{F}} W(x,\mathrm{F},z,\mathrm{A})\mathrm{F}^T$ is a "multiplicative stress" and is here estimated uniformly in terms of the energy density W. Assumption (2.2.22c) states uniform continuity of this stress tensor even if the energy density itself is considered as a weight – at least in a neighborhood of the identity.

Under the prerequisites which we have gathered until now, we can already formulate an existence result for solutions of the increment problems from which energetic solutions are constructed later on. These increment problems will be the focus of our algorithmic investigations in Chapters 3 and 4 which makes being aware of the assumptions needed for their well-definedness in the context of finite strain plasticity desirable.

Theorem 2.2.2: Existence of Solutions of Increment Problems

Consider $\mathcal{Q} \coloneqq \mathcal{Y} \times \mathcal{Z}$ according to (2.2.16) and (2.2.17) as well as the functionals $\mathcal{E}$ and $\mathcal{D}$ from (2.2.18). Assume that polyconvexity and coercivity from (2.2.20) hold for integrability powers p_{df}, p_{el}, p_{pl}, p_{hd} and p_{gr} sufficing (2.2.21). Moreover, assume that D satisfies (2.2.19) and (2.2.22) holds for W.

Then, for all partitions $\Pi \in \mathrm{Part}\big([0,T]\big)$, solutions to the correspondingly formulated time-incremental minimization problems $(\mathrm{IMP})^{\Pi}$ exist.

For the proof of the above theorem, we know from Proposition 2.1.8 that we require the abstract conditions (E1) and (E2) for the energy functional as well as (D1) and (D2) for the dissipation. We will explain how these can be deduced from their concrete counterparts formulated here towards the end of the current section.

Let us now continue with the analysis of energetic solutions of the time-dependent problem. For the formulation of the last assumption which we will need in order to guarantee the validity of the compatibility conditions (C1) and (C2), we have to introduce yet another class of functions. For $j, m_0, m_1, ..., m_j \in \mathbb{N}$, a mapping $a\colon \Omega \times \mathbb{R}^{m_1} \times ... \times \mathbb{R}^{m_j} \to \mathbb{R}^{m_0}$ is called a *Carathéodory function* if $a(\cdot, r_1, ..., r_j)\colon \Omega \to \mathbb{R}^{m_0}$ is measurable for all $(r_1, ..., r_j) \in \mathbb{R}^{m_1} \times ... \times \mathbb{R}^{m_j}$ and $a(x, \cdot)\colon \mathbb{R}^{m_1} \times ... \times \mathbb{R}^{m_j} \to \mathbb{R}^{m_0}$ is continuous for almost all $x \in \Omega$. Obviously, this notion can be extended to matrix-valued domains and we thus formulate

$$
\begin{aligned}
&D\colon \Omega \times \big(\mathrm{SL}(d) \times \mathbb{R}^m\big)^2 \to [0,\infty[\text{ is a Carathéodory function and} \\
&\exists\, h \in L^1(\Omega),\, C > 0,\, p_1 \in [1, p_{\mathrm{pl}}[,\, p_2 \in [1, p_{\mathrm{hd}}[\colon \\
&\quad \big|D(x, \mathrm{P}_0, \Pi_0, \mathrm{P}_1, \Pi_1)\big| \leq h(x) + C\big(\|\mathrm{P}_0\|^{p_1} + \|\mathrm{P}_0\|^{p_1} + \|\Pi_0\|^{p_2} + \|\Pi_0\|^{p_2}\big)
\end{aligned}
\tag{2.2.23}
$$

for all $x \in \Omega$ as well as $(\mathrm{P}_0, \Pi_0), (\mathrm{P}_1, \Pi_1) \in \mathrm{SL}(d) \times \mathbb{R}^m$. This assumption on the dissipation functional D is rather simple and restrictive concerning the applications in elasto-plasticity. Furthermore, it can be loosened to more general but also complicated formulations, cf. [66, Equation 4.2.11]. For our purposes of introducing *kinematic hardening*, however, this easily comprehensible variant of conditions leading to compatibility is sufficient.

With all of the above assumptions at hand, we are finally in the position to formulate the existence of energetic solutions in the framework of finite strain elasto-plasticity similar as in [66, Theorem 4.2.1]:

Theorem 2.2.3: Existence of Energetic Solutions

Consider $\mathcal{Q} \coloneqq \mathcal{Y} \times \mathcal{Z}$ according to (2.2.16) and (2.2.17) as well as the functionals $\mathcal{E}$ and $\mathcal{D}$ from (2.2.18). Assume that polyconvexity and coercivity from (2.2.20) together with the multiplicative stress bounds from (2.2.22) hold for integrability powers p_{df}, p_{el}, p_{pl}, p_{hd} and p_{gr} sufficing (2.2.21). Moreover, assume that D satisfies (2.2.19) and (2.2.23).

Then, for all stable initial conditions $\boldsymbol{q}_0 = (\boldsymbol{y}_0, \boldsymbol{z}_0)\colon \overline{\Omega} \to \mathbb{R}^d \times \big(\mathrm{SL}(d) \times \mathbb{R}^d\big)$, there exists an energetic solution $\boldsymbol{q}\colon [0,T] \to \mathcal{Q}$ for $\big(\mathcal{Q}, \mathcal{E}, \mathcal{D}, q_0\big)$.

Let us now shortly link together the assumptions formulated prior to the above existence result and explain how they are utilized in order to verify their rather abstract counterparts from Section 2.1. A detailed version of all arguments can be found in [66, Section 4.2.1].

As we have already pointed out beforehand, the polyconvexity from (2.2.20a) together with the regularity of the Dirichlet data driving the process lets us conclude weak lower semi-continuity under the additional assumption (2.2.21) on the integrability powers. Additionally, the coercivity from (2.2.20b) suffices to infer uniform boundedness of sublevels of $\mathcal{E}(t,\cdot)$ for all $t \in [0,T]$ by the adequate use of Hölder's inequality. These two steps then lead to the first demand on $\mathcal{E}$ itself, i.e., the assumption (E1).

Going on, we then use the uniform continuity of the derivatives via (2.2.22b) in addition to (2.2.20) and the regularity of the Dirichlet data in order to deduce in-time differentiability of the stored energy functional together with a closed-form expression for $\partial_t \mathcal{E}(t,\boldsymbol{q})$. From here, we use Gronwall's inequality for a uniform continuity estimate as in (2.2.22b) for the latter time-derivative. At last, a modulus of continuity for the gradients of the Dirichlet data helps us to conclude the energetic control of power as formulated in (E2).

The abstract requirements on the dissipation distance from (D1) and (D2) follow immediately from the assumptions we have made beforehand via (2.2.19).

As far as the compatibility conditions are concerned, an in-time uniform continuity result for $\partial_t \mathcal{E}$ on sublevels of $\mathcal{E}$ itself from the proof of (E2) together with (2.2.23) is used in order to provide the requirements of an auxiliary result stating sufficient assumptions for (C1) and (C2).

With all of the abstract assumptions on the corresponding initial value problem $\big(\mathcal{Q},\mathcal{E},\mathcal{D},q_0\big)$ at hand, the existence result in Theorem 2.2.3 for the elasto-plastic problem is a direct consequence of the more general result from Theorem 2.1.9.

2.2.4 Formulation of the Application Problem

In what follows, we will deduce the precise form of the $d = 3$-dimensional finite strain plasticity application problem which we will consider later on in Chapter 5 – at least up to the concrete object geometry and applied external forces. In particular, for a suitable initial state $(\boldsymbol{y}^0,\mathbf{P}^0)$ we will consider the *time-incremental minimization problems*

$$\begin{aligned}&\text{Find } \big(\boldsymbol{y}^1,\mathbf{P}^1\big),...,\big(\boldsymbol{y}^{N_\Pi},\mathbf{P}^{N_\Pi}\big) \text{ such that for } k\in\{1,...,N_\Pi\}:\\ &\big(\boldsymbol{y}^k,\mathbf{P}^k\big) \text{ minimizes } (\boldsymbol{y},\mathbf{P}) \mapsto \mathcal{E}\big(t_k,\boldsymbol{y},\mathbf{P}\big)+\mathcal{D}(\mathbf{P}^{k-1},\mathbf{P})\,.\end{aligned} \tag{2.2.24}$$

as first formulated in $(\mathrm{IMP})^\Pi$ in Section 2.1.2 and reformulate them such that our solution algorithm developed in Chapters 3 and 4 can be directly applied. All of these problems work with fixed time points t_k which is why even within all of the mappings marked as bold no explicit time-dependence is present.

The time-incremental minimization problems from (2.2.24) are sometimes also referred to as *homotopy step problems*. This designation stems from the idea that the discretization of time-dependent boundary forces and conditions in time has to be fine enough in order to resolve the *homotopy* of the underlying physical formulation.

Material Models and Hardening for the Stored Energy Functional

Our first deliberations concern the stored energy functional in the form of its corresponding densities which are then integrated over the domain $\Omega \subset \mathbb{R}^3$ of our test body. We remember the split into elastic and plastic energy densities plus external influences from (2.2.14) and

omit dependence of the material point $x \in \Omega$. Thus, we formulate $\mathcal{E}\colon [0,T] \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ as

$$\mathcal{E}(t, \boldsymbol{y}, \mathbf{P}) \coloneqq \int_\Omega W_{\mathrm{el}}\big(\mathbf{F}_{el}(x)\big) + W_{\mathrm{hd}}\big(\mathbf{P}(x)\big)\,\mathrm{d}x + \langle \boldsymbol{\ell}(t), \boldsymbol{y}\rangle$$

where the specific form of the external influences can be retraced in (2.2.15). For this split and $d = 3$, the polyconvexity assumption from (2.2.20a) reduces to the existence of some convex $\mathbb{W}\colon \mathbb{R}^{3\times3} \times \mathbb{R}^{3\times3} \times \mathbb{R} \to \mathbb{R}_\infty$ such that we can write

$$W_{\mathrm{el}}(\mathrm{F}) = \mathbb{W}\big(\mathrm{F}, \mathrm{cof}(\mathrm{F}), \det(\mathrm{F})\big)$$

for all $\mathrm{F} \in \mathrm{GL}^+(3)$. As elaborated on in [66, Section 4.1.1], popular such models additionally sufficing the corresponding coercivity bound $W_{\mathrm{el}}(\mathrm{F}) \geq c_0 + c_1\big\|\mathrm{F}\big\|_F^{p_{\mathrm{el}}}$ for the Frobenius norm $\big\|\cdot\big\|_F$ and some elastic growth rate $p_{\mathrm{el}} > 1$ are for instance the *Mooney-Rivlin* model $W_{\mathrm{MR}}\colon \mathrm{GL}^+(3) \to \mathbb{R}$ from [72, 88] given by

$$W_{\mathrm{MR}}(\mathrm{F}) \coloneqq a\big\|\mathrm{F}\big\|_F^2 + b\big\|\mathrm{cof}(\mathrm{F})\big\|_F^2 + c\det(\mathrm{F})^2 + V\big(\det(\mathrm{F})\big) \tag{2.2.25}$$

with constants $a, b, c > 0$ and some convex function $V\colon \mathbb{R} \to \mathbb{R}$ as well as the *Ogden* model $W_{\mathrm{Ogden}}\colon \mathrm{GL}^+(3) \to \mathbb{R}$ from [76] defined via

$$W_{\mathrm{Ogden}}(\mathrm{F}) \coloneqq \sum_{i=1}^{N} \alpha_i \mathrm{tr}\big((\mathrm{F}^T\mathrm{F})^{p_i}\big) + V\big(\det(\mathrm{F})\big) \tag{2.2.26}$$

with $N \in \mathbb{N}$, constant prefactors $\alpha_i > 0$, exponents $p_i \geq 1$ for $i \in \{1, ..., N\}$ and again some convex $V\colon \mathbb{R} \to \mathbb{R}$.

Another very popular model from engineering literature is the so-called *St. Venant-Kirchhoff* material model $W_{\mathrm{SVK}}\colon \mathrm{GL}^+(3) \to \mathbb{R}$ which can be seen as an extension of the geometrically linear elastic material model to the non-linear regime. It is defined via

$$W_{\mathrm{SVK}}(\mathrm{F}) \coloneqq \frac{1}{2}\mathbb{C}\mathrm{E}\colon \mathrm{E} = \frac{1}{2}\sum_{i,j,k,l=1}^{3} \mathbb{C}_{ijkl}\mathrm{E}_{ij}\mathrm{E}_{kl} \tag{2.2.27}$$

where $\mathrm{E} = \frac{1}{2}\big[\mathrm{F}^T\mathrm{F} - \mathrm{I}\big]$ is the Green-Lagrange strain tensor as it first appeared in (2.2.2) and $\mathbb{C} = \{\mathbb{C}_{ijkl}\}$ is the *fourth-order tensor of elastic moduli*. The latter is usually assumed to be positive definite and symmetric in the sense that it maps symmetric tensors to symmetric tensors, i.e., we demand[9]

$$\exists\, \alpha > 0\, \forall\, \mathrm{E} \in \mathbb{R}^{3\times3}_{\mathrm{sym}}\colon \mathbb{C}\mathrm{E}\colon \mathrm{E} \geq \alpha\big\|\mathrm{E}\big\|_F^2\,, \tag{2.2.28a}$$

$$\forall i, j, k, l \in \{1, 2, 3\}\colon \mathbb{C}_{ijkl} = \mathbb{C}_{jikl} = \mathbb{C}_{klij}\,. \tag{2.2.28b}$$

Assumption (2.2.28a) in particular yields that (2.2.27) is coercive of the form $W_{SVK}(\mathrm{F}) \geq \varepsilon_0\big\|\mathrm{F}\big\|_F^4 - \frac{1}{\varepsilon_0}$ for some $\varepsilon_0 > 0$. On the downside, however, the energy density is neither poly- nor quasiconvex[10]. Additionally, the Green-Lagrange tensor E is insensitive of the sign of F

[9] Here, $\mathbb{R}^{3\times3}_{\mathrm{sym}}$ refers to the respective subset of symmetric matrices.

[10] Quasiconvexity is a generalization of polyconvexity as introduced in (2.2.20a) and under suitable additional assumptions also yields weak lower semi-continuity of the stored energy functional, cf. [66, Proposition 4.1.5].

and thus is incapable of implementing the non-interpenetration condition $\det(\mathrm{F}) > 0$ which accordingly has to be treated rather as a separate constraint.

Material symmetry additionally reduces the number of independent constants in $\mathbb{C}$. In $d = 3$ and for so-called *isotropic*[11] media, there are only two degrees of freedom and the St. Venant-Kirchhoff model (2.2.27) reduces to

$$W_{\mathrm{SVK}}(\mathrm{F}) = \frac{1}{2}\lambda|\mathrm{tr}(\mathrm{E})|^2 + \mu\big\|\mathrm{E}\big\|_F^2 \tag{2.2.29}$$

where the constants λ and μ are called *Lamé coefficients* and $\mathrm{tr}(\mathrm{A})$ denotes the trace of the matrix A. The so-called *shear modulus* μ and the *bulk modulus* λ have to suffice $\mu > 0$ and $\lambda + \frac{2}{3}\mu > 0$ for the positive definiteness from (2.2.28a) to hold. The latter sum here is also referred to as the *modulus of compression*.

As far as the second part of the stored energy density, the hardening density W_{hd}, is concerned, we consider a simple *kinematic hardening term* of the form

$$W_{\mathrm{hd}}(\mathbf{P}) := k_1\big\|\mathbf{P}\big\|_F^{p_{\mathrm{pl}}} + k_2\big\|\nabla\mathbf{P}\big\|_F^{p_{\mathrm{gr}}} \tag{2.2.30}$$

with hardening parameters $k_1, k_2 > 0$, a plastic growth rate $p_{\mathrm{pl}} > 1$ and an additional regularizing growth rate $p_{\mathrm{gr}} > 1$. The Frobenius norm of the 3-tensor $\nabla\mathbf{P}$ is computed component-wise by $\big\|\nabla\mathbf{P}\big\|_F = \big(\sum_{i=1}^3 \big\|\mathrm{D}_{x_i}\mathbf{P}\big\|_F^2\big)^{\frac{1}{2}}$. Often, in particular in engineering literature, the case of no gradient regularization, i.e., $k_2 = 0$ in (2.2.30), is desired. Neglecting gradient terms within the hardening density and thereby also within the whole of the stored energy density W rules out the lower bound on the latter formulated in (2.2.20b). Thus, also the compactness of sublevel sets required in its abstract counterpart (E2) is not satisfied. The importance of this property for the existence of both solutions to the time-incremental minimization problems and energetic solutions in general has become apparent over the course of Section 2.1.3. Since we want to stay as close as possible to existence theory for our underlying finite strain plasticity problem, we will only consider computational examples with non-trivial gradient regularization within our numerical investigations in Chapter 5.

Discussion About Compatibility With Existence Theory

Let us now shortly remark on how the above models fit into the theoretical framework of existence of energetic solutions from Section 2.2.3: The main aspects of the requirements for existence which we take into consideration here are the coercivity parameters $p_{\mathrm{el}}, p_{\mathrm{pl}}, p_{\mathrm{gr}}$ from (2.2.20b) satisfying (2.2.21), and the polyconvexity of the elastic energy density as formulated in (2.2.20a). The value of p_{hd} is not of interest since we neglect additional internal variables within our model. Usually in literature, so-called *linear kinematic hardening* which is governed by quadratic norm terms, i.e., $p_{\mathrm{pl}} = p_{\mathrm{gr}} = 2$ in (2.2.30), is of interest. This does already rule out existence of solutions according to our theory. Even for larger exponents within hardening terms we encounter problems with models considered in engineering applications. The Mooney-Rivlin energy density from (2.2.25) is polyconvex and coercive but only allows for $p_{\mathrm{el}} = 2$ in (2.2.20b) which does not suffice. The also polyconvex Ogden material model from (2.2.26) is more flexible in the choice of exponents and the corresponding parameter p_{el} is determined by the minimal p_i within its definition. The accordingly high choice of both

[11] Isotropic media exhibit invariances under internal rotations. This means that the material "has no preferred direction".

the hardening and Ogden exponent would then again leave us with an extremely unphysical model which is irrelevant from an application-oriented standpoint. The last one of the popular models considered above, the St. Venant-Kirchhoff energy density from (2.2.27), on the other hand exhibits a relatively high $p_{\mathrm{el}} = 4$ and is cheap in computation but then again is ruled out due to lacking polyconvexity.

Altogether, we conclude that theory and application exhibit a conflict of interest as far as model parameters are concerned. The trade-off which we will take for the simulations conducted later on in Chapter 5 is that we will pursue physical significance by using the material models introduced above together with non-trivial gradient regularization.

Spinless Plastic Range and Dissipation Functional

With the complete description of the stored energy density and thereby functional at hand, we can turn our attention to the concrete form of the plastic range and the adequate definition of a corresponding dissipation functional. While in our section on general existence theory for energetic solutions we have worked with the canonical choice

$$\mathbf{P}(x,t) \in \mathrm{SL}(3) = \left\{ \mathrm{M} \in \mathbb{R}^{3\times 3} \,\middle|\, \det(\mathrm{M}) = 1 \right\}$$

for all $x \in \Omega$ and $t \in [0,T]$, we have mentioned already there that different definitions of the plastic range are possible and might be advantageous for the adequate formulation of particular finite strain plasticity problems. The above choice includes non-trivial both symmetric and non-symmetric parts of the plastic strain. Non-symmetric parts are also referred to as *rotational parts* of the corresponding matrix and can from a physical point of view be interpreted as the so-called *plastic spin*. This choice is very general but efficient implementation strategies struggle in rigorously aligning with existence theory. We will elaborate on these problems and possibilities to work around them later on.

For our formulation here, we will consider the case of *spinless plasticity* which – most importantly for us – allows both for moving within the bounds given by the assumptions from existence theory and for exploiting the ensuing problem structure for an efficient implementation of our solution algorithm from Chapters 3 and 4. Hence, we use

$$\mathbf{P}(x,t) \in \mathrm{SL}(3)^{+}_{\mathrm{sym}} = \left\{ \mathrm{M} \in \mathrm{SL}(3) \,\middle|\, \mathrm{M} \text{ is symmetric and positive definite} \right\}$$

for all $x \in \Omega$ and $t \in [0,T]$ as the definition of our plastic range. On this domain space for our plastic variable, we have to now define the *dissipation functional*, i.e., a distance functional which measures the energy that is dissipated when moving from one plastic state to another. As given in (2.2.18), this mapping is typically given as an integral

$$\mathcal{D}(\mathbf{P}_1, \mathbf{P}_2) \coloneqq \int_\Omega D\big(\mathbf{P}_1(x), \mathbf{P}_2(x)\big)\,\mathrm{d}x$$

over the test body domain $\Omega \subset \mathbb{R}^3$ with respect to the *dissipation density*

$$D\colon \mathrm{SL}(3)^{+}_{\mathrm{sym}} \times \mathrm{SL}(3)^{+}_{\mathrm{sym}} \to]-\infty, \infty]\,.$$

As we have mentioned beforehand, cf. (2.2.19), this density needs to fulfill the triangle inequality and $D(\mathrm{P}_1, \mathrm{P}_2) = 0$ must hold if and only if $\mathrm{P}_1 = \mathrm{P}_2$ holds within the plastic range.

However, the dissipation density does not have to be symmetric. We can construct distance densities with these properties as lengths of shortest *(weakly) differentiable paths*

$$\mathfrak{P}(\mathrm{P}_1,\mathrm{P}_2) \coloneqq \left\{\mathcal{P}\in W^{1,1}\big(0,1;\mathrm{SL}(3)^+_{\mathrm{sym}}\big)\,\middle|\,\mathcal{P}(0)=\mathrm{P}_1\,,\ \mathcal{P}(1)=\mathrm{P}_2\right\}$$

between two plastic states $\mathrm{P}_1,\mathrm{P}_2\in\mathrm{SL}(3)^+_{\mathrm{sym}}$. The corresponding lengths of these paths are then measured by the dissipation potential $\mathcal{R}\colon \mathrm{SL}(3)^+_{\mathrm{sym}}\times T\big(\mathrm{SL}(3)^+_{\mathrm{sym}}\big)\to]-\infty,\infty]$ on the tangent bundle $T\big(\mathrm{SL}(3)^+_{\mathrm{sym}}\big)$. For our plastic range, $\mathcal{R}(\mathrm{P},\dot{\mathrm{P}}) \coloneqq \big\|\mathrm{P}^{-1/2}\dot{\mathrm{P}}\mathrm{P}^{-1/2}\big\|_F$ is a possible definition which leads to the dissipation density

$$D(\mathrm{P}_1,\mathrm{P}_2) \coloneqq \inf\left\{\int_0^1 T_0\big\|\mathcal{P}(s)^{-\frac{1}{2}}\dot{\mathcal{P}}(s)\mathcal{P}(s)^{-\frac{1}{2}}\big\|_F\,\mathrm{d}s\,\middle|\,\mathcal{P}\in\mathfrak{P}(\mathrm{P}_1,\mathrm{P}_2)\right\}. \tag{2.2.31}$$

The scalar $T_0>0$ here represents the *yield stress* of the plastic deformation as it has already been considered in Section 2.2.2. This choice of the dissipation functional and ensuing density align with our formerly established existence theory and thus allows for the investigation of energetic solutions of the corresponding rate-independent system, cf. [66, Remark 4.2.9].

Due to the symmetry inherent to our choice of the plastic range $\mathrm{SL}(3)^+_{\mathrm{sym}}$, arguments from [63, Section 4] can be adopted in order to prove the following closed form expression of the dissipation density from (2.2.31), cf. [66, Remark 4.2.9]. The steps to be taken for this endeavor, however, dig deeply into the theory of the underlying manifold description of finite strain plasticity. A comprehensive version of the rigorous argumentation can be retraced in [43].

Lemma 2.2.4: Closed Form Expression for the Dissipation Density

Consider $P_1,P_2\in\mathrm{SL}(3)^+_{\mathrm{sym}}$. The dissipation distance as defined in (2.2.31) from P_1 to P_2 in $\mathrm{SL}(3)^+_{\mathrm{sym}}$ is explicitly given by

$$D(P_1,P_2) = T_0\left\|\log\left(\delta P^+_{\mathrm{sym}}\right)\right\|_F$$

where $\delta P^+_{\mathrm{sym}} \coloneqq P_1^{-\frac{1}{2}}P_2P_1^{-\frac{1}{2}}$ denotes the so-called *plastic increment* from P_1 to P_2.

In particular, the above closed form expression shows that the dissipation distance as defined in (2.2.31) from a plastic state P_1 to another P_2 only depends on the plastic increment $\delta P^+_{\mathrm{sym}} \coloneqq P_1^{-\frac{1}{2}}P_2P_1^{-\frac{1}{2}}$.

As a consequence, this allows us to reconstruct the endpoint P_2 of the transformation within the plastic range $\mathrm{SL}(3)^+_{\mathrm{sym}}$ by using the so-called *plastic update operator*

$$\Delta^+_{\mathrm{sym}}\colon \mathrm{SL}(3)^+_{\mathrm{sym}}\times\mathrm{SL}(3)^+_{\mathrm{sym}}\to\mathrm{SL}(3)^+_{\mathrm{sym}}\,,\quad \Delta(P,\delta P) \coloneqq P^{\frac{1}{2}}\delta P P^{\frac{1}{2}} \tag{2.2.32}$$

insofar that with the above definitions $P_2 = P_1^{\frac{1}{2}}\delta P^+_{\mathrm{sym}}P_1^{\frac{1}{2}} = \Delta^+_{\mathrm{sym}}(P_1,\delta P^+_{\mathrm{sym}})$ holds. In particular, the plastic update operator is a well-defined mapping into the plastic range, i.e., it preserves the symmetry and determinant of both arguments.

Reformulation of the Time-Incremental Minimization Problems

Since with the above formula for updating plastic states it suffices to know the initial state and the plastic increment, we can now use the latter in order to reformulate the time-incremental

minimization problems from (2.2.24) with the use of the plastic increment as a variable. In particular, this allows us to efficiently take advantage of the closed form expression of our dissipation distance from Lemma 2.2.4. Taking these first steps, we arrive at the problem of finding the pair of mappings $(\boldsymbol{y}^k, \delta\mathbf{P}^k) \in \mathcal{Y} \times \mathcal{Z}$ which minimizes

$$(\boldsymbol{y}, \delta\mathbf{P}^+_{\mathrm{sym}}) \mapsto \mathcal{E}\left(t_k, \boldsymbol{y}, \Delta^+_{\mathrm{sym}}\left(\mathbf{P}^{k-1}, \delta\mathbf{P}^+_{\mathrm{sym}}\right)\right) + T_0 \int_\Omega \left\|\log\left(\delta\mathbf{P}^+_{\mathrm{sym}}(x)\right)\right\|_F \mathrm{d}x \tag{2.2.33}$$

and afterwards setting $\mathbf{P}^k \coloneqq \Delta^+_{\mathrm{sym}}\big(\mathbf{P}^{k-1}, \delta\mathbf{P}^k\big)$ for all $k \in \{1, ..., N_\Pi\}$. Here, the closed form expression of D is already helpful but still rather inconvenient to compute due to the matrix-logarithm term. For this reason, we take advantage of yet another reformulation of the increment search problem which is enabled by the following lemma. Again, the symmetry assumption on our plastic range space $\mathrm{SL}(3)^+_{\mathrm{sym}}$ is very helpful here and the proof follows directly from the eigenvector-decomposition.

Lemma 2.2.5: Exponential Representation of $\mathrm{SL}(3)^+_{\mathrm{sym}}$-Matrices

For the domain space $\mathbb{S}^3_0 \coloneqq \left\{\mathrm{A} \in \mathbb{R}^{3\times3} \,\middle|\, \mathrm{A}^T = \mathrm{A}\,,\ \mathrm{tr}(\mathrm{A}) = 0\right\}$, the canonical matrix exponential $\exp : \mathbb{S}^3_0 \to \mathrm{SL}(3)^+_{\mathrm{sym}}$ is bijective.

This allows us to rewrite the time-incremental problem (2.2.33) in terms of the exponential representation of the global plastic increment $\delta\mathbf{P}^+_{\mathrm{sym}} = \exp(\delta\mathbf{B})$ for a uniquely determined function $\delta\mathbf{B}$ mapping $x \in \Omega$ to matrices from $\mathbb{S}^3_0$. These can also be understood as *tangential plastic increment mappings* since – from a manifold standpoint – the newly acquired search space is the tangent space to the former one at the identity. As a consequence, solving the time-incremental minimization problems from (2.2.24) is equivalent to finding a pair of mappings $(\boldsymbol{y}^k, \delta\mathbf{B}^k)$ which minimize the objective functional defined via

$$F(\boldsymbol{y}, \delta\mathbf{B}) \coloneqq \mathcal{E}\left(t_k, \boldsymbol{y}, \Delta^+_{\mathrm{sym}}\big(\mathbf{P}^{k-1}, \exp(\delta\mathbf{B})\big)\right) + T_0 \int_\Omega \left\|\delta\mathbf{B}\right\|_F \mathrm{d}x \tag{2.2.34}$$

and afterwards setting $\mathbf{P}^k \coloneqq \Delta^+_{\mathrm{sym}}\big(\mathbf{P}^{k-1}, \exp(\delta\mathbf{B}^k)\big)$ in order to obtain the subsequent plastic state for any $k \in \{1, ..., N_\Pi\}$.

Minimizing this objective functional features a handful of challenging aspects: Firstly, the energy functional $\mathcal{E}$ is in general non-convex and, secondly, the dissipation functional given by the scaled norm term is non-differentiable. Dealing with these peculiarities of the time-incremental minimization problems and finding an efficient way to solve them will be the central point of consideration within the rest of the present treatise.

Discussion about Incorporating Plastic Spin

Before taking on the challenge of designing an efficient solver for the finite strain plasticity homotopy step problems formulated above, we want to shortly discuss the incorporation of the aforementioned concept of plastic spin into our framework. As we have also pointed out already, this general approach requires a different definition of the plastic range as the whole of SL(3) and not only the subset of symmetric positive definite matrices. This choice for the plastic range also comes with a different definition of the dissipation potential $\mathcal{R}(\mathrm{P}, \dot{\mathrm{P}}) \coloneqq \left\|\dot{\mathrm{P}}\mathrm{P}^{-1}\right\|_F$ which leads to the dissipation density

$$D_{\mathrm{spin}}(\mathrm{P}_1,\mathrm{P}_2) \coloneqq \inf\left\{\int_0^1 T_0\big\|\dot{\mathcal{P}}(s)\mathcal{P}(s)^{-1}\big\|_F\,\mathrm{d}s \,\middle|\, \mathcal{P}\in\mathfrak{P}(\mathrm{P}_1,\mathrm{P}_2)\right\} \tag{2.2.35}$$

where now $\mathfrak{P}(\mathrm{P}_1,\mathrm{P}_2)$ denotes the set of (weakly) differentiable paths connecting P_1 and P_2 in SL(3). Again here, it can be shown that (2.2.35) only depends on the corresponding plastic increment $\delta\mathrm{P} \coloneqq P_2P_1^{-1}$ and the update operator from (2.2.32) can be defined correspondingly. All of these generalizations of our symmetric framework from above still align with the theory for existence of energetic solutions derived beforehand. Going on, however, problems regarding the reformulation of the ensuing homotopy step problems (2.2.24) start to appear.

In accordance with the arguments from [63, Section 4], the closed form expression via the log-formula from Lemma 2.2.4 for the above definition of the dissipation density with plastic spin requires the restriction of the Frobenius norm within the dissipation functional to the symmetric part of the plastic increment $\delta\mathrm{P}$. This restriction, however, excludes the ensuing definition of the dissipation functional from the theory for existence of energetic solutions due to the lacking quasi-distance property.

A straight-forward approach to resolving this peculiarity is to decompose the plastic increment into a symmetric part and a plastic spin part which cancels out in the log-formula. From there, the reduction to the $\mathbb{S}^3_0$ search space from above can be employed and the subproblem can again be formulated as the search for symmetric plastic increments. Then, however, the non-symmetric spin component has to be regarded in the update of the plastic state which lets two further problems arise: The first one is that, then, one has to demand isotropy in order to obtain a proper reformulation of the subproblem. The second and even greater problem is that, due to an inconvenient appearance of the product rule, the gradient of the current plastic state $\nabla\mathbf{P}$ can not be computed straight-forwardly. In order to then still obtain an efficient implementation of a solution method, one has to neglect gradient regularization which involves major problems with existence theory both for the energetic problem as a whole and the time-incremental problems. We have already elaborated on the importance of gradient regularization when it was introduced in (2.2.30).

However, also this problem with existence theory can be circumvented by arguing that for sufficiently small homotopy steps within the time discretization scheme, it can be assumed that also the plastic increments $\delta\mathrm{P}$ are very small in their Frobenius norm. From there, a straightforward computation shows that these small increments almost exclusively have a symmetric part which in general suggests the assumption of symmetric plastic increments for a sufficiently fine time discretization. As a consequence, Lemma 2.2.5 for the exponential representation of plastic increments can be used in order to reformulate the time-incremental minimization problems from (2.2.24) into a similar final form as we have established in (2.2.34).

A more diligent view onto the incorporation of plastic spin has been taken in [43] where both approaches have additionally been compared numerically and the strong connection between the formulations with and without plastic spin is investigated. As far as our considerations here are concerned, however, we only note that the concept of plastic spin itself and its necessity in modeling finite strain plasticity are subject to lively discussion, cf. [18]. Furthermore, the spinless approach, which we have pursued over the course of this section, suffices to significantly approve the functionality of our function space algorithm developed Chapters 3 and 4 for the solution of a demanding real-world problem. For this reason, we decided to stay on the side of rigorous existence theory of energetic solutions and well-definedness of the homotopy step problems from (2.2.24) without further assumptions in the spinless formulation.

Chapter 3

Second Order Semi-Smooth Proximal Newton Methods in Hilbert Spaces

As we have seen over the course of the last few sections, accurately modeling material behavior using rate-independent systems and afterwards formulating finite strain plasticity problems in function space is a very demanding endeavor. The goal now is to develop an efficient function space algorithm in order to tackle the resulting non-convex and non-smooth minimization problems in Hilbert spaces. We will devote the following two chapters to this task. In the current one, we will on the one hand lay out the general framework concerning underlying function spaces and assumptions on the objective functional and on the other hand focus on the introduction and functionality of our Proximal Newton method.

Chapter Outline

Our approach to the development of such an algorithm can be summarized as follows: At first, in Section 3.1 we will introduce basic notions of non-linear optimization, starting with simple iterative descent methods like Gradient and Newton approaches but incrementally softening the assumptions made on the smoothness of the objective function of the underlying minimization problem. Afterwards, we present our approach to Proximal Newton methods in Hilbert spaces in Section 3.2 with the emphasis on local fast convergence using exactly computed update steps and globalization via an additional norm term in the subproblem for step computation. As mentioned beforehand, the focus here rather lays on algorithmic functionality than efficiency.

Notational Remark

Let us first give a short notational remark: As before, by $n \in \mathbb{N}$ we denote the dimension of an underlying Euclidean space domain $\mathbb{R}^n$ endowed with some norm $\|\cdot\|$. For a differentiable mapping $f\colon \mathbb{R}^n \to \mathbb{R}$ we then write $f'(x)$ for its derivative at some $x \in \mathbb{R}^n$. This derivative is then an element of the dual space of $\mathbb{R}^n$, i.e., a linear functional. Whenever we need to refer to its primal counterpart, we do so by writing $\nabla f(x) \in \mathbb{R}^n$ for the gradient of f at x which suffices $\langle \nabla f(x), v\rangle_{\mathbb{R}^n} = \langle f'(x), v\rangle$ for all $v \in \mathbb{R}^n$. In general, we will often omit the typical notation of the dual pairing $\langle\cdot,\cdot\rangle$ for the sake of simplicity. In the same way, we distinguish second order derivatives $f''(x)$ (, i.e., bilinear forms,) from their matrix counterparts denoted by $\nabla^2 f(x)$. Later, in a more general Hilbert space setting, we will take even more care concerning the corresponding Riesz-Isomorphism $\mathfrak{R}$ between a Hilbert space X and its dual space X^*.

3.1 Basic Notions of Non-Linear Optimization

Before we will consider Proximal Newton methods in order to cope with non-smoothness and non-convexity within a composite optimization problem, we will provide the basis for formulating and elucidating the algorithm by introducing some basic notions of non-linear optimization. The goal is to first present rather simple and comprehensible base algorithms under a generous set of both differentiability and convexity assumptions in order to then augment the algorithmic ideas both in view of requirements for convergence (like smoothness and convexity) and methodical innovations (like damping, decrease criteria and inexactness). Due to the introductory focus of this section, we will rather explain concepts than give rigorous information about presented settings and methods, emphasizing central principles of non-linear optimization which we will come back to in the design of our algorithm in Section 3.2.

Section Outline

The section is structured straight-forwardly: In Section 3.1.1 we consider unconstrained, smooth optimization problems and present basic first and second order methods in order to find stationary points. Afterwards, Section 3.1.2 introduces the notion of semi-smoothness which constitutes an adequate framework for softening differentiability assumptions on the objective functional while still preserving advantageous convergence properties. At last, we set the stage for composite minimization problems featuring a non-smooth part of the objective functional in Section 3.1.3 and present Proximal Gradient and Newton methods in Euclidean space. Later on, we will modify the latter in order to deal with the finite strain plasticity problem presented beforehand.

3.1.1 Unconstrained Minimization Problems in Finite Dimensions

Let us start by considering the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{3.1.1}$$

where the mapping $f\colon \mathbb{R}^n \to \mathbb{R}$ is given as a continuously differentiable objective function. In general, the content of this section follows the elaborations of the introductory book [108].

In search of minimizers of f, we often draw back to optimality conditions. A necessary first-order optimality condition for $x_* \in \mathbb{R}^n$ to be a local minimizer of f is $f'(x_*) = 0$. For this reason, we call x_* a stationary point in this case:

Definition 3.1.1: Stationary Point

The element $x_* \in \mathbb{R}^n$ is called a **stationary point** of problem (3.1.1) if $f'(x_*) = 0$ holds.

General Descent Methods

In order to now find a stationary point $x_* \in \mathbb{R}^n$ of problem (3.1.1) we will consider so-called general descent methods generating a sequence of iterates $(x_k)_{k\in\mathbb{N}}$ along which the objective function value decreases. These methods can be structured as follows:

Algorithm 1: Model Algorithm for General Descent Methods

Data: Starting point $x_0 \in \mathbb{R}^n$
Initialization: $k = 0$;
while $f'(x_k) \neq 0$ **do**

1. Compute an admissible search direction $s_k \in \mathbb{R}^n$;
2. Compute an admissible step size $\sigma_k > 0$;
3. Update the current iterate $x_{k+1} \coloneqq x_k + \sigma_k s_k$;
4. Update the sequence index $k \to k + 1$.

end

There are two central algorithmic quantities which we have to specify in order to turn the above model algorithm into a concrete method for solving (3.1.1). The first one is the notion of *admissible search directions*. A first plausible property of search directions is that the slope of f in the considered direction is negative. To this end, we consider

$$0 \overset{!}{>} \lim_{t \to 0^+} \frac{f(x+ts) - f(x)}{\|ts\|} = \frac{f'(x)s}{\|s\|}$$

and thus call $s \in \mathbb{R}^n \setminus \{0\}$ a *descent direction* of the continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ in the point $x \in \mathbb{R}^n$ if $f'(x)s < 0$ holds. In order for s to then also be admissible for our method we have to impose a second requirement which enables global convergence of the procedure along adequately chosen subsequences:

Definition 3.1.2: Admissible Search Directions

A subsequence $(s_k)_{k\in\mathcal{K}}$ of **search directions** generated in Algorithm 1 is called **admissible** if the following two properties hold:

(i) $\forall k \in \mathcal{K}\colon f'(x_k)s_k < 0$ (, i.e., all s_k are *descent directions*).

(ii) $\left(\dfrac{f'(x_k)s_k}{\|s_k\|}\right)_{k\in\mathcal{K}} \to 0 \quad \Rightarrow \quad \big(f'(x_k)\big)_{k\in\mathcal{K}} \to 0\ .$

We will later on see that in the rather illustrative Euclidean space setting the norm $\|f'(x_k)\|$ characterizes the maximal slope.[12] This implies that item (ii) in the above definition asserts that if the slope of f along s_k converges to zero along the considered subsequence, so does the maximal slope.

The second algorithmic quantity to be determined is the step size σ which scales the descent direction in order to obtain the update for our sequence of iterates. Also here, we formulate a set of prerequisites which ensures a global convergence result of the ensuing descent method.

[12] In this section, we identify the norm on $\mathbb{R}^n$ with the one on its dual space and the one on the space of bilinear forms, i.e., we also write $\|f'(x)\|$ and $\|f''(x)\|$ without further specification.

Definition 3.1.3: Admissible Step Sizes

A subsequence $(\sigma_k)_{k\in\mathcal{K}}$ of **step sizes** generated in Algorithm 1 is called **admissible** if along the corresponding subsequence $(s_k)_{k\in\mathcal{K}}$ of search directions the following two properties hold:

(i) $\forall k \in \mathcal{K}\colon f(x_k + \sigma_k s_k) \leq f(x_k)$.

(ii) $\big(f(x_k + \sigma_k s_k) - f(x_k)\big)_{k\in\mathcal{K}} \to 0 \quad \Rightarrow \quad \left(\frac{f'(x_k)s_k}{\|s_k\|}\right)_{k\in\mathcal{K}} \to 0$.

This minimal set of prerequisites is in particular fulfilled by so-called *efficient* step sizes. Efficient step sizes σ for some descent direction s of f at x are characterized by

$$f(x + \sigma s) \leq f(x) - \gamma\left(\frac{f'(x)s}{\|s\|}\right)^2 \tag{3.1.2}$$

for some (in practice rather small) constant $\gamma > 0$. We will see later on that generalizations of this concept are very useful also for more involved minimization problems.

As mentioned beforehand, we have designed the above demands on search directions and step sizes such that the ensuing general descent method according to Algorithm 1 features global convergence results to a stationary point. We formulate this property of such procedures within the following theorem which can also be found in [108, Satz 8.7]:

Theorem 3.1.4: Global Convergence of General Descent Methods

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Assume that Algorithm 1 does not terminate after finitely many iterations and generates sequences (x_k), (s_k) and (σ_k). Consider an accumulation point x_* of (x_k) and let $(x_k)_{k\in\mathcal{K}}$ be the corresponding subsequence converging to x_* such that the sequences $(s_k)_{k\in\mathcal{K}}$ of search directions and $(\sigma_k)_{k\in\mathcal{K}}$ of step sizes are admissible. Then, x_* is a stationary point of f.

Even though the above specifications of general descent methods appear to be useful in the context, the question of how to compute admissible search directions and step sizes still remains open. We will close this gap within what follows now.

Gradient Methods

The most natural choice for search directions in Algorithm 1 are so-called *directions of steepest descent*. For some continuously differentiable f as above at x these directions are characterized by the minimization problem

$$d \coloneqq \operatorname*{arg\,min}_{\|\delta x\|=1} f'(x)\delta x\,. \tag{3.1.3}$$

Then, every element of the form $s = \lambda d$ for $\lambda > 0$ is referred to as a direction of steepest descent. It is easy to see that (as long as we consider the norm $\|x\| \coloneqq (x^T x)^{\frac{1}{2}}$) the minimization problem (3.1.3) has the unique solution $d = -\frac{\nabla f(x)}{\|f'(x)\|}$ which in particular determines directions of steepest descent as $s = -\lambda\nabla f(x)$ for any $\lambda > 0$.

So-called *Gradient methods* thus use the negative gradient $s_k \coloneqq -\nabla f(x_k)$ as a descent direction. The admissibility of negative gradients as search directions according to Definition 3.1.2 is apparent. The determination of admissible step sizes can then be implemented by

using the so-called *Armijo-rule*: For fixed parameters $\beta \in]0,1[$ and $\gamma \in]0,1[$ we thus choose the largest number $\sigma_k \in \{1, \beta, \beta^2, \dots\}$ such that

$$f(x_k + \sigma_k s_k) - f(x_k) \leq \sigma_k \gamma f'(x_k) s_k \tag{3.1.4}$$

holds along the search direction s_k. This procedure is well-defined and *finite* if s_k is a descent direction. Then, it yields an admissible sequence of step sizes according to Definition 3.1.3.

Another, rarely easy implementable but often analytically interesting choice for the step size σ is the so-called *minimization-rule* which determines the minimal value of the objective function along the given search direction. We can formalize this choice by defining

$$\sigma_k \coloneqq \operatorname*{arg\,min}_{\sigma \in [0,\infty[} f(x_k + \sigma s_k) \tag{3.1.5}$$

at every iterate x_k for the corresponding search direction s_k. We note here that this strategy is only well-defined under suitable convexity assumptions but then also provides us with an admissible sequence of step sizes.

As mentioned beforehand, the above choices can be implemented into the algorithmic framework of descent methods under suitable assumptions. Consequently, our global convergence result from Theorem 3.1.4 can be applied to both scenarios:

Corollary 3.1.5: Global Convergence of Gradient Methods

Accumulation points of the sequences of iterates generated by Gradient methods ($s_k \coloneqq -\nabla f(x_k)$) with step sizes chosen according to the Armijo-rule (3.1.4) or, with strong convexity of f, the minimization-rule (3.1.5) are stationary points of problem (3.1.1).

In addition to the global convergence behavior, local convergence rates of iterative minimization methods are often of interest:

Definition 3.1.6: Convergence Rates of Sequences

The sequence $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ converges ...

(i) ... q-linearly with rate $0 < \theta < 1$ to $\bar{x} \in \mathbb{R}^n$ if there exists some $k_0 \in \mathbb{N}$ such that

$$\forall k \geq k_0 \colon \|x_{k+1} - \bar{x}\| \leq \theta \|x_k - \bar{x}\| .$$

(ii) ... q-superlinearly to $\bar{x} \in \mathbb{R}^n$ if $x_k \to \bar{x}$ holds together with

$$\|x_{k+1} - \bar{x}\| = o(\|x_k - \bar{x}\|) \quad \text{for } k \to \infty .$$

This requirement is equivalent to the convergence of the fraction

$$\frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|} \to 0 \quad \text{for } k \to \infty .$$

(iii) ... q-quadratically to $\bar{x} \in \mathbb{R}^n$ if $x_k \to \bar{x}$ holds together with

$$\|x_{k+1} - \bar{x}\| = O(\|x_k - \bar{x}\|^2) \quad \text{for } k \to \infty .$$

This requirement is equivalent to the existence of some constant $C > 0$ and $k_0 \in \mathbb{N}$ such that

$$\forall k \geq k_0 \colon \|x_{k+1} - \bar{x}\| \leq C \|x_k - \bar{x}\|^2 .$$

Remark. *The above definitions can analogously be expanded to general metric spaces (X, d_X) by replacing norm difference terms in $\mathbb{R}^n$ by the corresponding metric expressions.*

Let us fill these concepts with life by considering the convergence rate of the Gradient method for a simple quadratic and strongly convex objective function $f\colon \mathbb{R}^n \to \mathbb{R}$, i.e., defined via

$$f(x) \coloneqq \langle c, x\rangle_{\mathbb{R}^n} + \frac{1}{2}\langle x, Cx\rangle_{\mathbb{R}^n} \quad , \, c \in \mathbb{R}^n \, , \, C \in \mathbb{R}^{n\times n} \text{ positive definite} \, .$$

In this case, we can show that even if the minimization-rule (3.1.5) for step size choice is employed, the corresponding Gradient method exhibits merely linear convergence with the rate depending on the maximal and minimal eigenvalues of the matrix C, cf. [108, Satz 7.8].

We can overcome this upper bound on the speed of convergence (at least near stationary points) for example by using second order information about our objective functional f which leads us to so-called *Newton methods*.

Newton Methods

Originally, *Newton's method* has been introduced as an iterative procedure in order to solve non-linear systems of equations of the form

$$F(x) = 0 \tag{3.1.6}$$

for some continuously differentiable $F\colon \mathbb{R}^n \to \mathbb{R}$. The iterative nature of the method can be explained as follows: For some given iterate $x_k \in \mathbb{R}^n$ we want to find an update step $s_k \in \mathbb{R}^n$ such that $F(x_k + s_k) = 0$ holds or at least the ensuing sequence of iterates approaches a solution of (3.1.6). To this end, we consider the Taylor-expansion of F via

$$F(x_k + s) = F(x_k) + F'(x_k)s + \rho(s)$$

where due to the continuous differentiability of F we have $|\rho(s)| = o\big(\|s\|\big)$ in the limit of $s \to 0$, i.e., that the remainder term is very small for small updates s. This suggests that computing the update s_k by solving the linearized system

$$F(x_k) + F'(x_k)s_k = 0 \tag{3.1.7}$$

yields a good approximation of the original problem $F(x_k+s_k) = 0$ at least if x_k is already close to a solution of (3.1.6). Before now considering both global and local convergence properties of the ensuing method for solving non-linear systems of equations, we will transfer the above idea to the problem of minimizing non-linear functions as in (3.1.1).

To this end, we now consider a twice continuously differentiable objective function $f\colon \mathbb{R}^n \to \mathbb{R}$. *Newton methods* for minimizing f can now be deduced coming from two different points of view. The first one is considering first order optimality conditions according to the definition of stationary points in Definition 3.1.1 and thereby solving

$$f'(x) = 0$$

where $f'\colon \mathbb{R}^n \to \mathbb{R}^n$ is still a continuously differentiable function governing a non-linear set of equations just as in (3.1.6). This yields the update system

$$f''(x_k)s_k = -f'(x_k) \tag{3.1.8}$$

to be solved in order to find an update step $s_k \in \mathbb{R}^n$.

The second approach to Newton methods for minimizing functions does not use its historical intention of solving non-linear systems of equations but directly considers the quadratic approximation of the objective function by the corresponding Taylor-expansion at some $x_k \in \mathbb{R}^n$ via

$$f(x_k + s) = f(x_k) + f'(x_k)s + \frac{1}{2}f''(x_k)(s)^2 + o\big(\|s\|\big)\,.$$

Similar as before, the idea is to now disregard the remainder term and only consider the quadratic model of f. Thus, we compute the update step via

$$s_k \coloneqq \operatorname*{arg\,min}_{s\in\mathbb{R}^n} q_k(s) \coloneqq f(x_k) + f'(x_k)s + \frac{1}{2}f''(x_k)(s)^2\,. \tag{3.1.9}$$

This minimization problem is only well-defined if the quadratic model q_k is convex, i.e., if the second order derivative $f''(x_k)$ is positive definite. Since we in particular want to investigate Newton methods close to minimizers of (3.1.1) and have continuous second order differentiability, we formulate the following *second order sufficient conditions* for strict local minimizers, cf. [108, Satz 5.5]:

Theorem 3.1.7: Sufficient Second Order Optimality Conditions

Consider the twice continuously differentiable function $f\colon U \to \mathbb{R}$ for some open set $U \subset \mathbb{R}^n$. A point $\bar{x} \in U$, which suffices the conditions

(i) $f'(\bar{x}) = 0$ (, i.e., $\bar{x}$ is a stationary point of (3.1.1)) and

(ii) $f''(\bar{x})$ is positive definite (, i.e., we have $\forall d \in \mathbb{R}^n \setminus \{0\}\colon f''(x)(d)^2 > 0$),

is a strict local minimizer of f.

As a consequence, the update rule (3.1.9) is well defined near (local) solutions of (3.1.1) which feature sufficient second order optimality conditions. The optimality of the ensuing update step s_k on the other hand can be characterized by

$$0 = q_k'(s_k) = f'(x_k) + f''(x_k)s_k$$

which is an equivalent reformulation of the update computation system (3.1.8). To conclude the above findings, we recognize that iteratively minimizing quadratic approximations of f and applying the original Newton's method for solving $f'(x) = 0$ lead to the same procedure for solving (3.1.1) when f is twice continuously differentiable.

With this update rule at hand, we can formulate the local version of Newton methods as introduced above for minimization problems:

Algorithm 2: Local Newton Method

Data: Starting point $x_0 \in \mathbb{R}^n$
Initialization: $k = 0$;
while $f'(x_k) \neq 0$ **do**

1. Compute the Newton update step s_k by solving $f''(x_k)s_k = -f'(x_k)$;
2. Update the current iterate $x_{k+1} \coloneqq x_k + s_k$;
3. Update the sequence index $k \to k+1$.

end

As pointed out beforehand, the information on the second order derivative of f utilized by Newton methods speeds up the local convergence of the ensuing algorithm. This expresses itself in the following form, cf. [108, Satz 10.8]:

Theorem 3.1.8: Local Convergence of the Local Newton Method for Minimization Problems

Consider $f\colon \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable with local minimum $x_* \in \mathbb{R}^n$ which features sufficient second order optimality conditions. Then, there exist $\delta > 0$ and $\mu > 0$ such that the following assertions hold:

(i) x_* is the only stationary point in $B_\delta(x_*)$.

(ii) Minimal eigenvalues suffice $\lambda_{\min}\big(f''(x)\big) \geq \mu$ for all $x \in B_\delta(x_*)$.

(iii) For starting points $x_0 \in B_\delta(x_*)$, Algorithm 2 either terminates with $x_k = x_*$ or generates a sequence $(x_k) \subset B_\delta(x_*)$ converging q-superlinearly to x_*.

(iv) If f'' is Lipschitz-continuous on $B_\delta(x_*)$, i.e., there exists $L > 0$ such that

$$\forall x, y \in B_\delta(x_*)\colon \big\|f''(x) - f''(y)\big\| \leq L\big\|x - y\big\|$$

holds, the convergence rate from (iii) is q-quadratic and we have

$$\forall k \in \mathbb{N}\colon \big\|x_{k+1} - x_*\big\| \leq \frac{L}{2\mu}\big\|x_k - x_*\big\|^2.$$

Obviously these local convergence rates are superior to the mere linear convergence of Gradient methods which we have investigated beforehand. However, note that they only consider the behavior close to minimizers of f which additionally feature sufficient second order optimality conditions. A global convergence result as in Theorem 3.1.5 can not be achieved for the simple version of Newton methods as in Algorithm 2. For some rather simple examples of f, divergence of the ensuing method can be shown, cf. [108, Section 10.3].

To this end, within the framework of Newton methods we introduce another important principle of non-linear optimization: globalization techniques. Globalization can be understood as the modification of iterative optimization methods such that regardless of the starting point x_0 a solution of the underlying minimization problem will be found by the algorithm. For the example of Newton methods, we still want to take advantage of the local fast convergence of

the base algorithm which suggests a modification of Algorithm 2 of the following form:

Algorithm 3: Globalized Newton Method

Data: Starting point $x_0 \in \mathbb{R}^n$, $\beta \in]0,1[$, $\gamma \in]0,1[$, $\alpha_1, \alpha_2 > 0$ and $p > 0$
Initialization: $k = 0$;
while $f'(x_k) \neq 0$ **do**

1. Compute a search direction s_k according to:
   If Solving $f''(x_k)d_k = -f'(x_k)$ is possible and the resulting d_k suffices

$$-f'(x_k)d_k \geq \min\left\{\alpha_1, \alpha_2 \|d_k\|^p\right\}\|d_k\|^2 :$$

   Then Set $s_k = d_k$;
   Else Set $s_k = -\nabla f(x_k)$;
2. Compute the step size σ_k according to the Armijo-rule (3.1.4);
3. Update the current iterate $x_{k+1} := x_k + \sigma_k s_k$;
4. Update the sequence index $k \to k+1$.

end

This algorithm uses (scaled) Newton update steps whenever possible and resorts to simple Gradient method updates whenever problems with either computability of the Newton steps occur or these Newton steps do not suffice the declared admissibility criterion. The latter indeed implies admissibility of the search direction s_k as introduced in Definition 3.1.2. A global convergence result for Algorithm 3 thus again follows directly by Theorem 3.1.4.

Our original goal, however, was to also take advantage of the local fast convergence of Algorithm 2. To this end, the *transition to local convergence* has to be guaranteed, i.e., that – at least sufficiently close to minimizers of (3.1.1) – the computation of Newton steps is always possible, yields a descent direction and the unit step size $\sigma_k = 1$ is accepted by the Armijo-rule (3.1.4). We summarize these findings within the following lemma, cf. [108, Lemma 10.13]:

Lemma 3.1.9: Transition to Local Convergence of the Globalized Newton Method

Consider $f\colon \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable and let x_* be a local minimizer of f featuring sufficient second order optimality conditions. Then, for any $\gamma \in]0, \frac{1}{2}[$ there exists $\delta > 0$ such that for any $x \in B_\delta(x_*)$ the following assertions hold:

(i) The vector $s := -\big(\nabla^2 f(x)\big)^{-1}\nabla f(x)$ is a descent direction of f at x.

(ii) The Armijo-condition is fulfilled for every $\sigma \in]0,1]$: $f(x+\sigma s) - f(x) \leq \sigma\gamma f'(x)s$.

This result now ensures that locally our globalized version of Newton's method from Algorithm 3 transitions into the locally accelerated procedure from Algorithm 2 and we can thus summarize its properties in the following theorem, cf. [108, Satz 10.14]:

Theorem 3.1.10: Convergence Properties of the Globalized Newton Method

Consider $f\colon \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable. Assume that Algorithm 3 generates a sequence (x_k) and let x_* be an accumulation point of this sequence with positive definite second order derivative $f''(x_*)$. Then, the following holds:

(i) The point x_* is an isolated local minimizer of f.

(ii) The whole sequence (x_k) converges to x_*.

(iii) There exists $k_0 \in \mathbb{N}$ such that for $k \geq k_0$ the algorithm adopts the update scheme of the Newton method from Algorithm 2. In particular, it exhibits q-superlinear convergence upgrading to q-quadratic convergence if f'' is Lipschitz-continuous in a neighborhood of x_*.

Summary of Concepts

Over the course of the above investigation of both (globalized) Newton and Gradient methods, we have gathered some key principles of non-linear optimization parallels to which we can also draw in more involved scenarios later on: Firstly, we can introduce *criteria for both search directions and step sizes* such that we can infer global convergence results for the ensuing iterative method. Secondly, *higher order information* on derivatives of the objective function can be very useful in order to gain better local convergence rates at least close to optimal solutions. Thirdly, we have to use *globalization strategies* in order to avoid divergence of iterative methods and ensure the successful minimization of the objective functional regardless of the starting point of the algorithm. Lastly, we have to guarantee *transition to local convergence* insofar that modifications from globalization techniques vanish as we approach solutions of the underlying minimization problem. This allows us to then take advantage of faster local convergence rates of higher order methods.

Inexact Newton Methods

Before we now head on to the task of softening smoothness assumptions on the objective functional of (3.1.1) away from second order continuous differentiability, we will introduce another slight modification of Newton methods as presented above. For large scale problems, the exact solution of the Newton update system (3.1.7) using direct methods like Gauß-elimination can be problematic or at least very expensive. Thus, iterative solvers for these subproblems are often used.

A natural question arising from this circumstance is whether terminating these inner solvers early and thereby only inexactly solving (3.1.7) still yields satisfying convergence results of the ensuing method. Assuming that a relative error estimate in the residual of the form

$$\big\|F(x_k) + F'(x_k)s_k\big\| \leq \eta_k \big\|F(x_k)\big\| \tag{3.1.10}$$

for some small forcing term $\eta_k \geq 0$ characterizes sufficient accuracy of the approximate solutions s_k of the subproblem, the following result about the so-called *inexact Newton's method* can be deduced, cf. [108, Satz 12.2]:

Theorem 3.1.11: Local Convergence of Inexact Newton Methods

Consider $F\colon \mathbb{R}^n \to \mathbb{R}^n$ to be continuously differentiable and x_* to be a root of F such that $F'(x_*)$ is invertible. Then, there exist $\delta > 0$ and $\eta > 0$ such that the following assertions are true:

(i) For some starting point $x_0 \in B_\delta(x_*)$ and update steps s_k sufficing (3.1.10) with $\eta_k < \eta$, the ensuing Inexact Newton Method either terminates with $x_k = x_*$ or generates a sequence (x_k) converging to x_* q-linearly.

(ii) If additionally $\eta_k \to 0$ holds, the convergence is q-superlinear.

(iii) If even more $\eta_k = O\big(\|F'(x_k)\|\big)$ holds and F' is Lipschitz-continuous on $B_\varepsilon(x_*)$, the convergence is q-quadratic.

This notion of inexact methods will be very useful for more involved minimization problems since they in particular feature harder to solve subproblems for finding update steps. Thus, inexactness in update step computation offers great potential of reducing computational effort while preserving advantageous convergence properties.

3.1.2 Semi-Smoothness and Semi-Smooth Newton Methods

Even though the above convergence results for corresponding minimization algorithms are satisfying in the context, the differentiability assumptions on the objective functions occur to be rather restrictive for many applications to be considered. For that reason, we will now introduce a generalized concept of smoothness and develop algorithms as well as convergence results for it, tailored to lacking second order continuous differentiability.

On this endeavor, we use generalized differentials in order to establish the notion of finite dimensional semi-smoothness and formulate semi-smooth Newton methods in the Euclidean $\mathbb{R}^n$. In order to extend the latter to Banach spaces, we first generalize the concept of semi-smoothness and afterwards adapt the algorithm to the infinite dimensionality of the new framework we find ourselves in. Our elaborations in general follow the ones in [111] where a both comprehensible and detailed introduction to the topic is given together with application of the concepts to demanding real-world problems.

Generalized Differentials and Finite Dimensional Semi-Smoothness

Let us consider a non-empty set $V \subset \mathbb{R}^n$ and a function $F\colon V \to \mathbb{R}^m$. By $D_F \subset V$ we denote the set of all $x \in V$ at which F admits a (Fréchet-)derivative $F'(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ represented by the matrix $\nabla F(x) \in \mathbb{R}^{m\times n}$ following the notational scheme from before. Now suppose that F is *locally Lipschitz-continuous*, i.e., for all $x \in V$ there exists an open neighborhood $U(x) \subset V$ together with a constant $L_x > 0$ such that for all $y \in U(x)$ we have the estimate:

$$\big\|F(x) - F(y)\big\| \leq L_x \big\|x - y\big\| .$$

Generalized notions of differentiability and the corresponding derivatives now evolve around the following classical theorem, cf. [120] or [28, Theorem 2.9.19]:

Theorem 3.1.12: Rademacher's Theorem

Suppose $F\colon V \to \mathbb{R}^m$ is locally Lipschitz continuous for $V \subset \mathbb{R}^n$. Then, F is differentiable almost everywhere on V.

In the above scenario, we in particular infer that $V \setminus D_F$ has Lebesgue measure zero. With this result at hand, the definition of the so-called *B-subdifferential* (B for Bouligand) is straightforward. It has first been considered in [80] and is given by

$$\partial_B F(x) \coloneqq \left\{ \mathrm{M} \in \mathbb{R}^{m\times n} \,\middle|\, \exists\, (x_k) \subset D_F \colon x_k \to x\,,\ F'(x_k) \to \mathrm{M} \right\}.$$

The corresponding convex hull

$$\partial_G F(x) \coloneqq \operatorname{co}\big(\partial_B F(x)\big) \coloneqq \left\{ \sum_{i=1}^{k} \lambda_i \mathrm{M}_i \,\middle|\, k \in \mathbb{N},\ \sum_{i=1}^{k} \lambda_i = 1,\ \mathrm{M}_i \in \partial_B F(x),\ 1 \le i \le k \right\}$$

as introduced in [15] is referred to as *Clarke's generalized Jacobian* of F at x. While some helpful properties of these generalized differentials can be verified, cf. [111, Proposition 2.2], we are rather interested in their role within the definition of at least finite dimensional semi-smoothness.

This has been done by Mifflin [68] for real-valued functions, and an extension to mappings between two finite dimensional spaces has then been given by Qi [85] as well as Qi and Sun [86]. The motivation for the concept is to develop locally q-superlinearly convergent Newton methods which are applicable despite the general non-smoothness of the underlying mapping. The following definition can be retraced to [68, 80, 86]:

Definition 3.1.13: Finite Dimensional Semi-Smoothness

Let $V \subset \mathbb{R}^n$ be non-empty and open. The function $F\colon V \to \mathbb{R}^m$ is called **semi-smooth** at $x \in V$ if it is Lipschitz continuous near x and if the following limit exists for all directions $s \in \mathbb{R}^n$:

$$\lim_{\substack{\mathrm{M} \in \partial_G F(x+\tau d) \\ d \to s\,,\ \tau \overset{>}{\to} 0}} \mathrm{M}d\,.$$

If F is semi-smooth at all $x \in V$, we call F **semi-smooth (on V)**.

This definition of semi-smoothness does not obviously exhibit the properties which we demanded by our motivation for the concept. However, there are several connections between directional differentiability and semi-smoothness as introduced above. For instance, the limit from Definition 3.1.13 coincides with the corresponding directional derivative if both exist. Additionally, we have the following characterizations of semi-smoothness which use directional differentiability, cf. [111, Proposition 2.7]:

Proposition 3.1.14: Characterizations of Finite Dim. Semi-Smoothness

Consider an open set $V \subset \mathbb{R}^n$ and the function $F\colon V \to \mathbb{R}^m$. Then, for $x \in V$, the following statements are equivalent:

(a) F is semi-smooth at x.

(b) F is Lipschitz continuous near x, $F'(x,\cdot)$ exists, and

$$\sup_{\mathrm{M}\in\partial_G F(x+s)} \big\|\mathrm{M}s - F'(x,s)\big\| = o\big(\|s\|\big) \text{ in the limit of } s \to 0\,.$$

(c) F is Lipschitz continuous near x, $F'(x,\cdot)$ exists, and

$$\sup_{\mathrm{M}\in\partial_G F(x+s)} \big\|F(x+s) - F(x) - \mathrm{M}s\big\| = o\big(\|s\|\big) \text{ in the limit of } s \to 0\,. \quad (3.1.11)$$

In particular, continuously differentiable functions are semi-smooth.

This shows that semi-smoothness as introduced in Definition 3.1.13 is equivalent to Lipschitz continuity and directional differentiability together with an approximation property of the corresponding elements of Clarke's generalized Jacobian $\partial_G F$. These Clarke derivatives either yield a direct approximation of the directional derivative or a first order model of F itself. The latter property of semi-smooth functions will in particular be useful for the analysis of Newton-type methods.

In Proposition 3.1.14, we have already mentioned that continuously differentiable functions are in particular semi-smooth. Other examples for semi-smooth mappings are convex real-valued functions, continuous and piecewise C^1-functions, locally Lipschitz and tame (, in particular, semi-algebraic) functions[13], eigenvalues of symmetric matrices, singular values of general matrices as well as spectral operators induced by semi-smooth functions, cf. [112].

With the finite dimensional notion of semi-smoothness at hand, we can now formulate and analyze a Newton-like method for the solution of the equation (3.1.6) where now $F\colon V \to \mathbb{R}^n$ for an open $V \subset \mathbb{R}^n$ is semi-smooth at the solution.

Algorithm 4: Finite Dimensional Local Semi-Smooth Newton Method

Data: Starting point $x_0 \in \mathbb{R}^n$
Initialization: $k = 0$;
while $F(x_k) \neq 0$ **do**

1. Choose $\mathrm{M}_k \in \partial_G F(x_k)$ and compute s_k from $\mathrm{M}_k s_k = -F(x_k)$;
2. Update the current iterate $x_{k+1} \coloneqq x_k + s_k$;
3. Update the sequence index $k \to k+1$.

end

Under a suitable regularity assumption on the matrices M_k, $k \in \mathbb{N}$, this iteration converges locally q-superlinearly. We formulate this result within the following proposition, cf. [111, Proposition 2.12]:

[13] For a definition of tame and semi-algebraic functions as well as a proof of their semi-smoothness, we refer to [9].

Proposition 3.1.15: Local Convergence of the Finite Dimensional Semi-Smooth Newton Method

For an open set $V \subset \mathbb{R}^n$ and a mapping $F\colon V \to \mathbb{R}^n$, consider an isolated solution $x_* \in \mathbb{R}^n$ of (3.1.6). Assume the following:

(a) The approximation property (3.1.11) holds at $x = x_*$ (which, in particular, is satisfied if f is semi-smooth at x_*).

(b) There exists a constant $C > 0$ such that for all $k \in \mathbb{N}$ the matrices M_k are non-singular with the bound $\left\|\mathrm{M}_k^{-1}\right\| \leq C$.

Then, there exists $\delta > 0$ such that for all $x_0 \in B_\delta(x_0)$ Algorithm 4 either terminates with $x_k = x_*$ or generates a sequence (x_k) that converges q-superlinearly to x_*.

The regularity assumption (b) from Proposition 3.1.15 can be formulated as a general assumption on F at the solution x_* (so-called *CD-regularity*) instead of the above estimate for the inverse of M_k at the iterates x_k, cf. [111]. In [52], this requirement has been adapted to a uniform injectivity condition (CI) which corresponds to the boundedness of inverse matrices in norm here.

Generalization of Semi-Smoothness to Banach Spaces

As we have learned over the course of Chapter 2, many real-world applications are modeled using infinite dimensional function spaces in which the corresponding minimization problems then have to be solved by suitable solution algorithms. Thus, the natural question arises whether semi-smooth Newton methods like Algorithm 4 can be developed in an infinite dimensional framework. In particular, this question is not only of theoretical interest but also of practical importance since the performance of numerical methods is intimately related to the infinite dimensionality of the underlying problem. In particular, rather than in developing discrete numerical methods for discretized problems, our interest lies in investigating a well-behaved abstract algorithm for the infinite dimensional problem itself. The convergence analysis of the latter is then able to predict the convergence properties of the numerical algorithm very well, even for increasing accuracy of the discretization. Consequently, the adequate approach to the development of robust, efficient, and mesh-independent numerical methods is the investigation of corresponding algorithms within the original infinite dimensional problem setting. The publications [41, 109, 110] have provided a rigorous basis for the later intensively investigated and successfully applied semi-smooth Newton methods in function spaces.

As we have seen in the formulation of Proposition 3.1.15, the crucial point of semi-smoothness for the analysis of Newton-type methods is the approximation property (3.1.11). Using this estimate, we recall the following abstract concept of semi-smoothness for general operators between Banach spaces from [111].

Definition 3.1.16: Semi-Smooth Operators in Banach Spaces

Consider Banach spaces X and Y together with an open subset $V \subset X$ and a mapping $F\colon V \to Y$. Further, let a set-valued mapping $\partial^* F\colon V \rightrightarrows \mathcal{L}(X, Y)$ be given with non-empty images, i.e., $\partial^* F(x) \neq \emptyset$ for all $x \in V$.

(a) We say that F is **$\partial^* F$-semi-smooth** at $x \in V$ if F is continuous near x and the

following approximation property holds:

$$\sup_{\mathrm{M}\in\partial^* F(x+s)} \big\|F(x+s) - F(x) - \mathrm{M}s\big\|_Y = o\big(\|s\|_X\big) \text{ in the limit of } \|s\|_X \to 0\,.$$

(b) We will refer to $\partial^* F\colon V \rightrightarrows \mathcal{L}(X,Y)$ as the **generalized differential** of F, and the non-emptiness of its images will always be assumed. In particular, the $\partial^* F$-semi-smoothness of F at a point $x \in V$ shall automatically imply that the images of $\partial^* F$ are non-empty on V.

While the specific choice of $\partial^* F$ depends on the application at hand, in general it can be interpreted as a set-valued point-based approximation. For further elaborations, consider [52, 89, 116]. Basic properties of this infinite dimensional understanding of the concept like semi-smoothness of continuously differentiable functions with respect to their (Fréchet-) derivative, semi-smoothness of sums and direct products as well as a chain rule are developed in [111, Section 3.2.2] and provide a variety of ways to combine semi-smooth operators in order to construct new ones.

In this context, also the notion of *Newton-differentiability* and corresponding *Newton-derivatives* often arises, cf. e.g. [16]. The definition of the former is closely connected to our above generalized notion of semi-smoothness: A continuous mapping $F\colon V \to Y$, $V \subset X$ open, between Banach Spaces X and Y is thus called *Newton-differentiable* at $x \in X$ if there exists a neighborhood $N(x) \subset X$ and a mapping $G\colon N(x) \to \mathcal{L}(X,Y)$ with

$$\big\|F(x+s) - F(x) - \mathrm{G}(x+s)s\big\|_Y = o\big(\|s\|_X\big) \text{ in the limit of } \|s\|_X \to 0\,.$$

Any mapping $M \in \big\{G(\tilde{x}) \,\big|\, \tilde{x} \in N(x)\big\}$ is then called a *Newton-derivative* of F at x. However, taking a closer look at that definition reveals a major flaw of this general concept of Newton-differentiability: The mapping G characterizing Newton-derivatives can always be defined depending on the base point x which makes Newton-differentiability a trivial property of the underlying mapping F. For the application of the notion to minimization strategies, concepts of differentiability are of particular interest at optimal solutions which are not known a priori which makes the above definition even more questionable. Additionally, every mapping $M \in \mathcal{L}(X,Y)$ is a Newton-derivative in the above sense since the characterizing property of G only narrows these down in the limit $s \to 0$.

One way to resolve this peculiarity of describing admissibility of a mapping for Newton-like approaches to minimization has been described in [41] where the notion of *slant-differentiability* has been used in order to study semi-smooth Newton methods. Slant-differentiability modifies the above understanding of Newton-differentiability insofar that it is not formulated *in a certain point* $x \in X$ but on a whole subset $U \subset D$ of the corresponding function domain. This eliminates the possibility to simply tailor the corresponding mapping G to the point where differentiability is investigated. Transferred to the context of Newton-methods, slant-differentiability "parallels the hypothesis of knowledge of the domain within which a second order sufficient optimality condition is satisfied for smooth problems", cf. [41, Page 3].

Our approach towards the problem, however, is rather coined by the following thought: More than in the existence of a mapping G and a corresponding concept of differentiability, we are interested in whether the characterizing approximation property of such a mapping is satisfied by a mapping which is given, e.g. within the framework of a minimization procedure. Thus, the mapping in question has to be part of the definition and not only the existence of

a mapping of such kind. This motivates the following notion which both works around the above peculiarities of general Newton-differentiability and simplifies general semi-smoothness insofar that we do not have to define a generalized differential. In particular, this slightly different formulation is not considered in [111].

Definition 3.1.17: Semi-Smoothness with Respect to an Operator

Consider Banach spaces X, Y, an open subset $V \subset X$, a point $x \in V$, and a neighborhood $N(x) \subset V$ of x. The continuous mapping $F\colon V \to Y$ is called **semi-smooth** at x **with respect to the operator** $G\colon N(x) \to \mathcal{L}(X,Y)$ if the following approximation property holds:

$$\big\|F(x+s) - F(x) - \mathrm{G}(x+s)s\big\|_Y = o\big(\|s\|_X\big) \text{ in the limit of } \|s\|_X \to 0\,.$$

We then call G a **Newton-derivative** of F at $x \in V$.

Obviously, this notion is closely related to the one of $\partial^* F$-semi-smoothness from Definition 3.1.16: If some continuous mapping F is $\partial^* F$-semi-smooth at $x \in V$, the generalized differential gives a plethora of possible mappings G with respect to which F is then semi-smooth. At every point $\tilde{x}$ of, in this case, $N(x) = V$, we can choose $G(\tilde{x}) \in \partial^* F(\tilde{x})$ arbitrarily. Conversely, with a mapping $G\colon N(x) \to \mathcal{L}(X,Y)$ at hand, we define $\partial^* F$ arbitrarily outside of the designated neighborhood $N(x)$ and use the single-valued definition $\partial^* F(\tilde{x}) = \{G(\tilde{x})\}$ for $\tilde{x} \in N(x)$ which then again yields $\partial^* F$-semi-smoothness of F.

Semi-Smooth Newton Methods in Banach Spaces

In addition to the adaptations of the concept of semi-smoothness to the infinite dimensionality of the underlying problem, also the algorithmic strategy from Algorithm 4 itself has to be augmented in order to cope with function space applications of the method. Let us consider the following short motivational example from [111] for this adjustment:

A common field of application for semi-smooth Newton-like methods are *Non-linear Complementarity Problems (NCPs)* of the form

$$u \geq 0\,, \quad \Gamma(u) \geq 0\,, \quad u\Gamma(u) = 0 \quad \text{on } \Omega$$

with $u \in L^2(\Omega)$, an operator $\Gamma\colon L^2(\Omega) \to L^2(\Omega)$, and $\Omega \subset \mathbb{R}^n$ a bounded, measurable set with positive Lebesgue measure. This problem can be reformulated with a very particular choice of the *NCP-function* $\phi\colon \mathbb{R}^2 \to \mathbb{R}$ such that we obtain

$$\Phi(u) = 0\,, \quad \text{where} \quad \Phi(u)(\omega) = \phi\big(u(\omega), F(u)(\omega)\big) \quad \text{for all } \omega \in \Omega \tag{3.1.12}$$

with $F\colon L^p(\Omega) \to L^{p'}(\Omega)$ for $p, p' \in]1, \infty]$, and $\Phi\colon L^2(\Omega) \to L^2(\Omega)$ the latter of which needs to be semi-smooth for our solution algorithm. In a quite general setting, however, we are merely able to show that $\Phi\colon L^p(\Omega) \to L^2(\Omega)$ is semi-smooth for $p > 2$. Embedding the latter result into a semi-smooth Newton context, for convergence analysis it would be necessary to assume that the operators $\mathrm{M}_k \in \partial^* \Phi(u_k)$ are boundedly invertible in $\mathcal{L}\big(L^p(\Omega), L^2(\Omega)\big)$, which is usually not satisfied.

What resolves this technical peculiarity is the incorporation of so-called *smoothing steps* into our algorithmic framework, cf. [111]. This enables us to work in a setting where, given the availability of a suitable smoothing step, we only require the semi-smoothness of $\Phi\colon L^p(\Omega) \to$

$L^2(\Omega)$ for some $p > 2$ together with the bounded invertibility of M_k in $\mathcal{L}\big(L^2(\Omega), L^2(\Omega)\big)$ which appear to be both appropriate and verifiable in the context.

The construction and particular structure of such smoothing steps goes beyond the scope of the present work but can be retraced for challenging applications in [109, Section 6] or [111, Section 3.3]. In order to formulate the concept in an abstract setting, we consider a further Banach space X_0 (above $L^2(\Omega)$), in which X (above $L^p(\Omega)$) is continuously and densely embedded. Thus, we introduce the following semi-smooth Newton method for operator equations in Banach spaces:

Algorithm 5: Local Semi-Smooth Newton Method in a Banach Space X

Data: Starting point $x_0 \in \mathbb{R}^n$
Initialization: $k = 0$;
while $x_{k+1} \neq x_k$ **do**

1. Choose $\mathrm{M}_k \in \partial^* F(x_k)$, compute s_k from $\mathrm{M}_k s_k = -F(x_k)$
2. Set the intermediate iterate $x_{k+1}^0 \coloneqq x_k + s_k.$;
3. Perform a smoothing step: $X_0 \ni x_{k+1}^0 \mapsto x_{k+1} \coloneqq S_k(x_{k+1}^0) \in X$;
4. Update the sequence index $k \to k + 1$.

end

Even though we have already elaborated on the significance of the smoothing step in applications, it can be eliminated from the above algorithmic procedure by simply choosing $X_0 = X$ and $S_k(x_{k+1}^0) = x_{k+1}^0$. In general, we will assume the corresponding non-trivial step to satisfy the so-called *smoothing condition* which we formulate by the existence of a uniform constant $C_S > 0$ such that

$$\big\|S_k(x_{k+1}^0) - x_*\big\|_X \leq C_S \big\|x_{k+1}^0 - x_*\big\|_{X_0} \tag{3.1.13}$$

holds for all $k \in \mathbb{N}$, where $x_* \in X$ solves (3.1.6). If we now as before additionally demand the *regularity condition*, stating that the operators M_k map X_0 continuously into Y and that there exists a uniform constant $C_{M^{-1}} > 0$ such that we have

$$\big\|M_k^{-1}\big\|_{\mathcal{L}(Y,X_0)} \leq C_{M^{-1}} \tag{3.1.14}$$

for all $k \in \mathbb{N}$, we can establish the following local convergence result for Algorithm 5, cf. [111, Theorem 3.13]:

Theorem 3.1.18: Local Convergence of the Infinite Dimensional Semi-Smooth Newton Method

Consider Banach spaces X, Y and an open set $V \subset X$ as well as a mapping $F\colon V \to Y$ with generalized differential $\partial^* F\colon V \rightrightarrows \mathcal{L}(X, Y)$. Suppose that $x_* \in V$ is a solution of (3.1.6) and that assumptions (3.1.13) and (3.1.14) hold. If then F is $\partial^* F$-semi-smooth at x_*, there exists $\delta > 0$ such that for all $x_0 \in B_\delta(x_*)$ Algorithm 5 either terminates with $x_k = x_*$ or generates a sequence $(x_k) \subset V$ that converges q-superlinearly to x_*.

Having the above algorithmic framework and corresponding local convergence result in place, there are still two crucial questions left to address before we move on:

(a) Given a particular operator F, how should $\partial^* F$ be chosen?

(b) Is there an easy way to verify that F is $\partial^* F$ semi-smooth?

Giving satisfactory answers to these questions goes beyond the scope of this introductory section for the concept of semi-smoothness. Generally, the answer to (a) depends on the application at hand and, obviously, (b) is intimately tied to this specific choice of the generalized differential. In our motivational example for the smoothing step in (3.1.12), we could already get a glimpse of a central topic for the solution of non-smooth operator equations: so-called *superposition operators*. In short, they can be understood as mappings of the form

$$\Psi\colon X \to L^r(\Omega)\,, \quad \Psi(x)(\omega) \coloneqq \psi\big(G(x)(\omega)\big)$$

with mappings $\psi\colon \mathbb{R}^m \to \mathbb{R}$ and $G\colon X \to \Pi_{i=1}^m L^{r_i}(\Omega)$ where it is assumed that $1 \leq r \leq r_i < \infty$ holds, X is a real Banach space, and $\Omega \subset \mathbb{R}^m$ is a bounded measurable set with positive Lebesgue measure. In [111, Section 3.3] both of the above questions are answered for this particular kind of mappings and the corresponding results are illustrated for the NCP example from (3.1.12). An alternative approach to showing semi-smoothness of superposition operators has been given in [97], which resulted in a more comprehensible proof and better understanding of the concept as a whole. We will go more into detail on continuity results about superposition operators when investigating *second order semi-smoothness* in Section 3.2.4.

Before heading on to incorporating actual non-differentiability into our optimization problem, let us also here shortly reflect on the main concepts we have learned about over the course of this introductory section: Firstly, the notion of *semi-smoothness* both in an Euclidean and in an infinite dimensional Banach space setting allows us to soften differentiability assumptions insofar that only an approximation property of corresponding *generalized derivatives* is necessary for the formulation of locally superlinearly convergent minimization algorithms. Secondly, the transition from semi-smoothness in $\mathbb{R}^n$ to its infinite dimensional counterpart gave insights into the development of generalized notions of differentiability as a whole: Instead of sticking to existing concepts like in the finite dimensional case, in the more general setting it suffices to use the *approximation property* which comprises the spirit of semi-smoothness, in particular in view of its application for the construction of the corresponding Newton methods. We will encounter a similar thought process within the development of *second order semi-smoothness* later on in this chapter.

3.1.3 Proximal Methods for Composite Optimization

In addition to the softened differentiability assumptions in the form of semi-smoothness from the previous section, we want to now incorporate actual non-differentiability into our formulation, i.e., we will consider problems of the form

$$\min_{x \in \mathbb{R}^n} F(x) \coloneqq f(x) + g(x) \tag{3.1.15}$$

which is for now posed in the finite dimensional $\mathbb{R}^n$ with a somewhat "smooth" $f\colon \mathbb{R}^n \to \mathbb{R}$ and "non-smooth" $g\colon \mathbb{R}^n \to]-\infty, \infty]$. Special cases for these kind of problems among other examples comprise the unconstrained *smooth minimization problems* from the previous sections for $g = 0$, *constrained minimization* for $g = \mathcal{X}_C$ with C some (convex) set to which x in (3.1.15) is restricted, and so-called l_1*-regularized minimization* for $g = \lambda\|\cdot\|_1$ a scaled 1-norm term

which promotes sparsity within solutions of the respective minimization problem. While all of these instances feature a rather "artificial" non-smooth term, a non-trivial function g also can arise from a natural formulation like in the case of our time-incremental minimization problems from (2.2.34) within the finite strain plasticity problem.

Proximal Gradient Methods

In order to now develop an iterative solution algorithm for problems of the form (3.1.15), we will take a look at an equivalent way to compute the Gradient method update $x_{k+1} = x_k - \sigma_k \nabla f(x_k)$ for some appropriately chosen step size $\sigma_k > 0$ from Section 3.1.1. The updated iterate from there can be rephrased by formulating the *subproblem of the Gradient method*

$$x_{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2\sigma_k} \big\| x - x_k \big\|^2$$

which can also be interpreted as minimizing a regularized first order model of f based at the current iterate $x_k \in \mathbb{R}^n$. Apparently, if we want to add a non-differentiable part g into this formulation, we cannot use a Taylor series of some form but have to directly incorporate it into the subproblem via

$$x_{k+1} \coloneqq \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2\sigma_k} \big\| x - x_k \big\|^2 + g(x) \,. \tag{3.1.16}$$

For the previously mentioned example of constrained optimization ($g = \mathcal{X}_C$) this leads to an update scheme coinciding with the so-called *Projected (Sub-)Gradient method*, cf. e.g. [7, Section 8.2]. Simple algebraic manipulation provides us with an alternative form of (3.1.16) given by

$$x_{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; \sigma_k g(x) + \frac{1}{2} \big\| x - \big(x_k - \sigma_k \nabla f(x_k) \big) \big\|^2 \,.$$

This identity can be written in terms of the so-called *Euclidean Proximal operator* of some function $g \colon \mathbb{R}^n \to]-\infty, \infty]$ which we define as

$$\operatorname{prox}_g \colon \mathbb{R}^n \to \mathbb{R}^n \,, \quad \operatorname{prox}_g(y) \coloneqq \operatorname*{arg\,min}_{x \in \mathbb{R}^n} g(x) + \frac{1}{2} \big\| x - y \big\|^2 \tag{3.1.17}$$

such that we obtain $x_{k+1} = \operatorname{prox}_{\sigma_k g}\big(x_k - \sigma_k \nabla f(x_k) \big)$ for the updated iterate. This update scheme justifies the designation of the ensuing procedure as the *Proximal Gradient method* since a gradient step is followed up by the evaluation of a proximal mapping for the completion of the update. Additionally, this formulation suggests the somewhat vague assumption that g features an "easy to evalute" prox-operator. This is, for example, the case if g and also the employed scalar product have diagonal structure. Then, the solution of the subproblem within the proximity operator can be computed cheaply in a componentwise fashion. For more elaborations on finite dimensional proximal operators, general properties, and closed form expressions for common examples, we refer to [7, Chapter 6].

Originally, Fukushima and Mine introduced the Proximal Gradient method in the Euclidean $\mathbb{R}^n$ for optimization problems of the above form (3.1.15), cf. [31]. More specifically, this early version of the Proximal Gradient method constitutes a special case of a procedure studied by Tseng and Yun, cf. [107]. Further research showed that variously defined line search

techniques lead to global convergence of the algorithm even under appropriate inexactness conditions for the solutions of the subproblem for step computation, cf. for example [11, 29, 32, 53, 96, 102]. For an illustration of the functionality and for giving an intuition about first convergence results, we follow the elaborations from the introductory book [7, Chapter 10]:

To this end, we consider the following set of assumptions for the composite objective functional: Let $g\colon \mathbb{R}^n \to]-\infty,\infty]$ be proper, closed, and convex. Let $f\colon \mathbb{R}^n \to]-\infty,\infty]$ be proper and closed, $\operatorname{dom}(f)$ be convex with $\operatorname{dom}(g) \subset \operatorname{int}(\operatorname{dom}(f))$ where $\operatorname{int}(A)$ denotes the interior of some set $A \subset \mathbb{R}^n$. Additionally, assume that f is continuously differentiable with Lipschitz-continuous derivative and that the set of solutions of (3.1.15) is non-empty.

For the sake of an easier analysis, it is often useful to consider the step size σ from above in terms of the fraction $\frac{1}{L}$ for some parameter $L > 0$ and with that define the so-called *Euclidean Composite Gradient mapping*

$$G_L^{f,g}\colon \mathbb{R}^n \to \mathbb{R}^n\,, \quad G_L^{f,g}(x) \coloneqq L\Big(x - \operatorname{prox}_{\frac{1}{L}g}\big(x - \frac{1}{L}\nabla f(x)\big)\Big) \tag{3.1.18}$$

which we will later on also generalize to a Hilbert space scenario. This mapping coincides with the classical gradient of f for $g = 0$, vanishes if and only if x is a critical point of (3.1.15), and allows us to rephrase the Proximal Gradient update from above via

$$x_{k+1} = x_k - \frac{1}{L_k} G_{L_k}^{f,g}(x_k)$$

where $L_k > 0$ is the value of L within the corresponding iteration. This representation of the update also mirrors the behavior of the classical Gradient method from Section 3.1.1.

Furthermore, the above adaptation of the notion of derivatives not only helps us to characterize critical points of our problem but also gives an intuition on how to find adequate values of the regularization parameter L in order to achieve sufficient decrease. The first possibility is rather of analytical nature and suggests choosing $L \in]\frac{L_f}{2}, \infty[$ constantly where L_f denotes the Lipschitz constant of f'. From an algorithmic point of view, it seems reasonable to transfer the procedure of the Armijo rule from (3.1.4) to the present Proximal Gradient scenario in the form of the following *Backtracking Procedure*, cf. [7, Section 10.3.3]:

The strategy requires three parameters (s, γ, ζ), where $s > 0$, $\gamma \in]0,1[$, and $\zeta > 1$. Then, we choose L_k according to the following scheme: At first, we set L_k to be equal to the initial guess s. Then, as long as $F(x_{k+1}) - F(x_k) > -\frac{\gamma}{L_k}\big\|G_{L_k}^{f,g}(x_k)\big\|^2$ holds, we increase the parameter via $L_k \coloneqq \zeta L_k$. The resulting value can also be rephrased as $L_k = s\zeta^{i_k}$, where $i_k \in \mathbb{N}$ is the smallest exponent for which the following *sufficient decrease criterion* is satisfied:

$$F(x_{k+1}) - F(x_k) \leq -\frac{\gamma}{s\zeta^{i_k}}\big\|G_{s\zeta^{i_k}}^{f,g}(x_k)\big\|^2\,. \tag{3.1.19}$$

Here, it is important to note that the above backtracking procedure is again finite, i.e., that (3.1.19) is fulfilled after finitely often increasing L_k within one backtracking loop. This results in the bound $L_k \leq \max\{s, \frac{\zeta L_f}{2(1-\gamma)}\}$ which is in particular uniform concerning the iteration index $k \in \mathbb{N}$, i.e., over the course of the Proximal Gradient method as a whole.

With this algorithmic component at hand, we can summarize our first minimization strategy for (3.1.15) as follows:

Algorithm 6: Euclidean Proximal Gradient Method

Data: Starting point $x_0 \in \mathbb{R}^n$, backtracking parameters $s > 0$, $\gamma \in]0,1[$, and $\zeta > 1$;
Initialization: $k = 0$;
while $G_s^{f,g}(x_k) \neq 0$ **do**

1. Choose L_k according to the backtracking procedure with parameters $s > 0$ $\gamma \in]0,1[$, and $\zeta > 1$;
2. Compute the updated iterate via $x_{k+1} = \operatorname{prox}_{\frac{1}{L_k} g}\big(x_k - \frac{1}{L_k}\nabla f(x_k)\big)$;
3. Update the sequence index $k \to k+1$.

end

The globalization via the backtracking procedure leading to a sufficient decrease estimate (3.1.19) allows us to verify the following convergence results for the algorithmic strategy from Algorithm 6, cf. [7, Theorem 10.15]. The definition of *stationary points* is carried over from the smooth case in Definition 3.1.1 using the Fréchet-subdifferential $\partial_F F$ from (1.3.3).

Theorem 3.1.19: Convergence of the Euclidean Proximal Gradient Method (Non-Convex)

Under the standing assumptions on f and g, let $(x_k)_{k\in\mathbb{N}}$ denote the sequence generated by Algorithm 6. Then, the following assertions hold:

(i) The sequence $\big(F(x_k)\big)_{k\in\mathbb{N}}$ is non-increasing. In addition, we have $F(x_{k+1}) < F(x_k)$ if and only if x_k is not a stationary point of (3.1.15).

(ii) The gradient mapping $G_s^{f,g}(x_k)$ converges to zero for $k \to \infty$.

(iii) All accumulation points of the sequence $(x_k)_{k\in\mathbb{N}}$ are stationary points of (3.1.15).

As it is emphasized within the caption of the above convergence result, this formulation does not assume convexity of the smooth part f of the composite objective functional. An additional convexity assumption on said mapping is sufficient to prove better convergence estimates, in particular an $O(1/k)$ decrease rate within function values as well as within the (minimal) norm of gradient mappings along our sequence of iterates, cf. [7, Theorems 10.21 and 10.26].

This additional convexity assumption does not only help us to verify better convergence results for the existing minimization method above but also allows for an algorithmic augmentation which further boosts convergence. The idea is to follow a *momentum-based* approach, i.e., going even further in the direction of the current Proximal Gradient update with an appropriately defined scaling factor stemming from the convergence analysis. This method is known as the "fast iterative shrinkage-thresholding algorithm" (FISTA) which can be summarized using the following scheme:

Algorithm 7: Accelerated Proximal Gradient Method – FISTA

Data: Starting point $x_0 \in \mathbb{R}^n$, backtracking parameters $s > 0$, $\gamma \in]0,1[$, and $\zeta > 1$;
Initialization: $k = 0$, $y_0 = x_0$, $t_0 = 1$;
while $G_s^{f,g}(x_k) \neq 0$ **do**

1. Choose L_k according to the backtracking procedure with parameters $s > 0$ $\gamma \in]0,1[$, and $\zeta > 1$ (as well as y_k instead of x_k);
2. Compute the updated iterate via $x_{k+1} = \operatorname{prox}_{\frac{1}{L_k}g}\big(y_k - \frac{1}{L_k}\nabla f(y_k)\big)$;
3. Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
4. Compute $y_{k+1} = x_{k+1} + \big(\frac{t_k - 1}{t_{k+1}}\big)(x_{k+1} - x_k)$;
5. Update the sequence index $k \to k+1$.

end

This augmentation of the Proximal Gradient method from Algorithm 6 unlocks a $O(1/k^2)$ rate in function values despite the fact that the dominant computational steps at each iteration of both methods are essentially the same: one gradient evaluation and one prox computation.

Proximal Newton Methods

As we have seen over the course of our investigation of Newton methods in Section 3.1.1, there is another possibility to accelerate the (at least local) convergence of minimization procedures: taking advantage of second order information in case the objective is sufficiently differentiable. In our case, the assumption of sufficient smoothness only is related to the smooth part f and the resulting algorithm is then referred to as the *Proximal Newton method*. Since it is our goal to elaborately develop a generally applicable and refined instance from this algorithmic class, we will only give a short formal introduction here:

The second order information enters the computational scheme in the form of a second order model of f instead of the first order one in (3.1.16). Additionally, the norm term is generally omitted which results in the definition of a search direction via

$$\Delta x_k \coloneqq \operatorname*{arg\,min}_{d \in \mathbb{R}^n} \hat{f}_k(x_k + d) - f(x_k) + g(x_k + d) - g(x_k) \tag{3.1.20}$$

where $\hat{f}_k \colon \mathbb{R}^n \to]-\infty, \infty]$, $\hat{f}_k(x) \coloneqq f(x_k) + f'(x_k)(x - x_k) + f''(x_k)(x - x_k)^2$, denotes said second order model of f based at the current iterate $x_k \in \mathbb{R}^n$. The ensuing update step is then appropriately scaled by some step size $\sigma_k > 0$ in order to obtain the new iterate according to $x_{k+1} \coloneqq x_k + \sigma_k \Delta x_k$. Global convergence results are then, as before, achieved by an adequate choice of σ_k such that some sufficient decrease criterion in the spirit of the Armijo rule (3.1.4) holds. This formulation is in particular used in [55] which gives an illustrative overview of Proximal Newton methods together with some adaptations in the Euclidean $\mathbb{R}^n$. From there, we also take the basic results of the introductory elaborations here.

In this framework, for the update step computation subproblem from (3.1.20) to be well-defined, uniform positive definiteness of the second order derivatives $f''(x_k)$ together with convexity of g is demanded. In the strongly convex case and if f'' is additionally Lipschitz-continuous, it is thus easy to see that the ensuing method still exhibits the local quadratic convergence typical for Newton-type procedures.

In many applications of the method, it seems reasonable to allow for approximations of the possibly not positive definite $f''(x_k)$ in the form of general bilinear forms H_k having the demanded positive-definiteness property. As long as these approximations suffice the so-called *Dennis-Moré condition*

$$\frac{\big\|\big(H_k - f''(x_*)\big)(x_{k+1} - x_k)\big\|}{\big\|x_{k+1} - x_k\big\|} \to 0 \quad \text{for } k \to \infty \tag{3.1.21}$$

at some optimal solution $x_* \in \mathbb{R}^n$ of (3.1.15), the ensuing *Proximal Quasi-Newton method* still exhibits at least local superlinear convergence.

Another algorithmic principle which we have already addressed for general Newton methods is the one of inexactness in update step computation. In the Euclidean space setting which we find ourselves in here, it is reasonable to generalize the inexactness criterion from (3.1.10) with the aid of the composite gradient mapping defined in (3.1.18) and thus formulate

$$\big\|G_M^{\hat{f}_k,g}(x_k + \Delta s_k)\big\| \leq \eta_k \big\|G_M^{f,g}(x_k)\big\| \tag{3.1.22}$$

as a requirement for the inexact candidate Δs_k where M denotes a uniform upper bound on the eigenvalues of the bilinear forms H_k used for update step computation. By $\hat{f}_k$, we refer to a second order model of f at the current iterate $x_k \in X$ and the so-called *forcing term* $\eta_k \geq 0$ controls the extent of inexactness within step computation. The criterion can be understood as a minimal decrease requirement for the gradient mapping which signifies sufficient "quality" of the corresponding update. Note that for the evaluation of (3.1.22), knowledge of M is essential which is not accessible in general applications. Additionally, we will later on see why an inexactness criterion of this form is not adequate for a function space scenario and the inherent infinite dimensionality of the corresponding formulation.

In the Euclidean $\mathbb{R}^n$, however, (3.1.22) allows to prove that – in spite of inexact computation of update steps – local linear convergence of the ensuing method is achieved if the forcing terms η_k stay below some upper bound. If the latter additionally converge to zero along the sequence of iterates, even local superlinear convergence can be verified. For the sake of completeness, let us write down the algorithmic procedure described above:

Algorithm 8: Euclidean Inexact Proximal Quasi-Newton Method

Data: Starting point $x_0 \in \mathbb{R}^n$, initial forcing term η_0;
Initialization: $k = 0$;
while $G_M^{f,g}(x_k) \neq 0$ **do**

1. Choose second order bilinear forms $(H_k)_{k\in\mathbb{N}}$ with (3.1.21);
2. Choose forcing term $\eta_k \geq 0$;
3. Compute search direction Δs_k sufficing (3.1.22) for η_k via (3.1.20) with H_k;
4. Choose step size σ_k according to sufficient decrease criterion;
5. Set $x_{k+1} = x_k + \sigma_k \Delta s_k$;
6. Update the sequence index $k \to k + 1$.

end

The convergence results which we have gathered beforehand will now be summarized within the following theorem which constitutes a selection of assertions from [55]:

Theorem 3.1.20: Convergence Properties of the Inexact Proximal Quasi-Newton Method

Consider the minimization problem (3.1.15) where $f, g\colon \mathbb{R}^n \to]-\infty, \infty]$ are proper, convex and f is twice continuously differentiable with Lipschitz second order derivative. Additionally, let f be strongly convex at the unique optimal solution of (3.1.15).

Then, the sequence of iterates computed by Algorithm 8 converges globally to that optimal solution. Locally, the unit step length $\sigma_k = 1$ is accepted and the convergence rate is ...

(i) ... q-linear if $\eta_k < \tilde{\eta} < \frac{m}{2}$ for m a lower bound on the eigenvalues of the H_k.

(ii) ... q-superlinear if $\eta_k \to 0$ for $k \to \infty$.

(iii) ... q-quadratic if the second order bilinear forms are chosen as $H_k = f''(x_k)$.

Also within this development of a suitable minimization strategy, we recognize the algorithmic principles which we have summarized towards the end of Section 3.1.1: *Globalization* by sufficient decrease criteria, *local acceleration* by second order information, *transition to local convergence* by admissibility of the unit step length, and *inexactness* which reduces computational effort but preserves convergence properties in qualified situations. We will keep these principles in mind for later.

Even though sufficient for introductory purposes, the above formulation still exhibits some major flaws: Convexity and differentiability assumptions are much too restrictive for demanding applications and the Euclidean domains do not allow for the treatment of function space problems.

Before dealing with these problematic properties of existing Proximal Newton methods, let us give a short overview, how problems of the form (3.1.15) are approached in the literature:

Obviously, further assumptions on the form of the composite objective functional open the door to more specific adaptions of the solution algorithm. For example in [22, 56, 104], the authors assume convexity and self-concordance[14] of the smooth part f in order to employ damped Proximal Newton methods. Alternatively, reformulations of the original minimization problem can be useful. As a consequence, methods which have been proven to work for other problem classes can also be applied in our case. For instance, in [12, 13, 57] fixed point algorithms are employed, or consider [4] for a reformulation of (3.2.1) as a constrained problem which opens up a whole field of different suitable solution approaches.

A different point of view onto this class of problems was taken by Milzarek and Ulbrich in [70]. For $g(x) \coloneqq \lambda\|x\|_1$ with $\lambda > 0$, they consider a semi-smooth Newton method with filter globalization which Milzarek later on generalizes to work also for arbitrary convex functions for g, cf. [69].

Rather recently, Kanzow and Lechner have discussed a globalized, inexact and possibly non-convex Proximal Newton-type method in Euclidean space $\mathbb{R}^n$, cf. [44]. There, the algorithm resorted to Proximal Gradient steps in the case of insufficient descent together with a line-search procedure in order to achieve global convergence and cope with lacking convexity of the objective functional.

The restrictive convexity assumptions from above can also be taken care of by trust region methods in order to make the step computation subproblem well-defined in spite of lacking convexity of the corresponding objective. This approach is pursued in [3], where the authors thoroughly study the respective Quasi-Newton method and present a rich convergence analysis.

[14] A function $f \in \mathrm{C}^3(\mathbb{R};\mathbb{R})$ is called *self-concordant* if $|f'''(x)| \leq 2f''(x)^{3/2}$ holds for all $x \in \mathbb{R}$. Self-concordant functions within Proximal Newton methods can be used in order to describe and modify the so-called *covariance selection problem* from [20], see [56].

3.2 Exact Computation of Update Steps

Even though many of the aforementioned works have accomplished significant progress with respect to formulating Proximal Newton methods for a wide range of applications, we will generalize the notion of this type of methods even further in order to allow for the solution of non-convex, non-smooth variational problems in function space like the finite strain plasticity problem presented in Chapter 2. To this end, we will add innovative modifications to the established algorithmic ideas in order to evade the existing rather restrictive assumptions concerning both the underlying vector-spaces and objective functions to be minimized. In that regard, our goal is to address three major deficiencies of present methods: Firstly, they are often presented in a finite dimensional framework prohibiting the application to function space problems. Secondly, convexity of the objective functions is a common assumption which we can not guarantee in general. Lastly, we will also replace classical differentiability with adequate notions of semi-smoothness which still yield desirable convergence results for our algorithm. The content of this section is closely related to the published paper [84].

Let us depart on this endeavor by reminiscing the generally considered optimization problem which now reads

$$\min_{x\in X} F(x) \coloneqq f(x) + g(x) \tag{3.2.1}$$

where $f : X \to \mathbb{R}$ is assumed to be smooth in some adequate sense and $g\colon X \to]-\infty, \infty]$ is possibly not. The domains of both f and g are given by subsets of an arbitrary real Hilbert space $(X, \langle\cdot,\cdot\rangle_X)$ with corresponding norm $\|v\|_X = \sqrt{\langle v, v\rangle_X}$ and dual space X^*. The Hilbert space structure of X also gives us access to the Riesz-Isomorphism $\mathfrak{R}\colon X \to X^*$, defined by $\mathfrak{R}x = \langle x, \cdot\rangle_X$, which satisfies $\|\mathfrak{R}x\|_{X^*} = \|x\|_X$ for every $x \in X$. Since $\mathfrak{R}$ is non-trivial in general, we will not identify X and X^*.

Assumptions on the Smooth Part

We will assume the smooth part of our objective functional $f\colon X \to \mathbb{R}$ to be continuously differentiable on its domain with Lipschitz-continuous derivative $f'\colon X \to X^*$, i.e., we can find some constant $L_f > 0$ such that for every $x, y \in X$ the estimate

$$\|f'(x) - f'(y)\|_{X^*} \leq L_f \|x - y\|_X \tag{A1}$$

holds.[15]

Similar as in the introductory section for (Proximal) Newton methods, we will refer to the quadratic part within the second order model of f at some $x \in X$ as $H_x \in \mathcal{L}(X, X^*)$. Furthermore, we will notationally identify the linear operators $H_x \in \mathcal{L}(X, X^*)$ with the corresponding symmetric bilinear form $H_x\colon X \times X \to \mathbb{R}$, and write $(H_x v)(w) = H_x(v, w)$, using the abbreviation $H_x(v)^2 \coloneqq H_x(v, v)$. We will assume uniform boundedness of H_x along the sequence (x_k) of iterates by the existence of some uniform $M > 0$ such that

$$\forall\, k \in \mathbb{N}\colon \|H_{x_k}\|_{\mathcal{L}(X,X^*)} \leq M \tag{A2}$$

holds. In addition, we assume the existence of a mapping $\kappa_1\colon X \to \mathbb{R}$ bounded from below such that along the sequence of iterates (x_k) we have the bound

[15] In what follows, referring to (A1) in particular implicitly demands continuous differentiability of f.

$$\forall\, k \in \mathbb{N}\ \forall v \in X\colon H_{x_k}(v)^2 \geq \kappa_1(x_k)\big\|v\big\|_X^2 \tag{A3}$$

which can be interpreted as an ellipticity assumption on H_{x_k} if $\kappa_1(x_k)$ is positive. In this case, when considering exact (and smooth) Proximal Newton methods, where H_x is given by the Hessian of f at some point $x_k \in X$, (A3) is equivalent to $\kappa_1(x_k)$-strong convexity of f at x_k. When considering general bilinear forms H without dependence on some x_k, we will refer to (A3) in the sense of $H(v)^2 \geq \kappa_1$ for some constant $\kappa_1 \in \mathbb{R}$ and all $v \in X$.

While in a sufficiently smooth setting $H_x \coloneqq f''(x)$ is common, for most of the paper we may choose H_x freely in the above framework. For fast local convergence, however, we will impose a semi-smoothness assumption in the form of an approximation property according to its introduction in Section 3.1.2, see (3.2.11) for the explicit formulation within the current scenario. Furthermore, in order to guarantee transition of our globalization scheme to fast local convergence, we suppose f to suffice the notion of *second order semi-smoothness* (cf. Section 3.2.4) with respect to the mapping having the above H_x as images. This concept generalizes second order differentiability in our setting and its definition is directly distinguishable from the one of semi-smoothness of f' in (3.2.11).

Assumptions on the Non-Smooth Part

We assume that the non-smooth part $g\colon X \to]-\infty, \infty]$ is proper, lower semi-continuous, and satisfies a bound of the form

$$g(sx + (1-s)y) \leq sg(x) + (1-s)g(y) - \frac{\kappa_2}{2}s(1-s)\big\|x-y\big\|_X^2 \tag{A4}$$

for some uniform $\kappa_2 \in \mathbb{R}$ where $x, y \in X$ and $s \in [0,1]$ can be chosen arbitrarily.[16] For $\kappa_2 > 0$ estimate (A4) represents κ_2-strong convexity of g. It is known that κ_2-strong convexity of g implies that g is bounded from below, its level-sets $L_\alpha g$ are bounded for all $\alpha \in \mathbb{R}$, and their diameter shrinks to 0 if $\alpha \to \inf_{x\in X} g$. In the case of $\kappa_2 < 0$, g is allowed to be non-convex in a limited way. This scenario is sometimes also referred to as $(-\kappa_2)$-*weak convexity* of g stating that $g - \frac{\kappa_2}{2}\big\|\cdot\big\|_X^2$ is convex for that $\kappa_2 < 0$.

As we have pointed out within the assumption on g in the introductory Section 3.1.3, an important practical aspect of splitting methods, such as Proximal Newton, is that the non-smooth part g of the composite objective functional F yields a proximal operator prox_g that can be evaluated easily. In function space problems, in particular if Sobolev spaces are involved, it is known that instead of a diagonal structure for the underlying scalar product, a multi-level structure should be used in order to reflect the topology of the function space properly. For this reason, diagonal proximal operators would suffer from mesh-dependent condition numbers. In our numerical computations, we therefore employ non-smooth multigrid techniques to compute the Proximal Newton steps, in particular *Truncated Non-smooth Newton MultiGrid methods* (TNNMG). For a more detailed description of the method, we refer to Section A.1 in the appendix, and for convergence results and applications to other (standalone) problems, consider [33, 92].

The theory behind Proximal Newton methods and the respective convergence properties evolves around the convexity estimates stated in (A3) and (A4). We will assign particular

[16] Also here, when demanding (A4) we implicitly include the assumption of g being proper and lower semi-continuous.

importance to the interplay of the convexity properties of f and g, i.e., the sum $\kappa_1(x)+\kappa_2$ will continue to play an important part over the course of the current chapter. Estimates (A1)-(A4) constitute the standing assumptions on our composite objective functional and implicitly hold in what follows even though they might not always be listed explicitly.

Section Outline

With the above assumptions on the underlying space and objective functional at hand, the rest of the section is structured as follows: At first, in Section 3.2.1, we will introduce a helpful tool for the convergence analysis of our method, the so-called *Dual Proximal Mappings*, and elaborate on some of their key properties. Next, in Section 3.2.2, we will consider undamped update steps computed as the solution of an adequately formulated subproblem and prove local superlinear convergence of the ensuing method. Afterwards, in Section 3.2.3, we present a modification of the aforementioned subproblem in order to damp update steps and globalize the Proximal Newton method. This enables the proof of optimality for all limit points of the sequence of iterates generated by our method. Section 3.2.4 then concerns the introduction of second order semi-smoothness for f and showcases how it helps to verify the admissibility of both full and damped update steps sufficiently close to optimal solutions in Section 3.2.5. This in turn enables the transition to local fast convergence of our globalized method. Lacking numerical robustness of the globalization strategy is addressed and improved by the introduction of an alternative sufficient decrease criterion in Section 3.2.6. In Section 3.2.7, the performance of this stage of the algorithm is substantiated by numerical results considering a rather simple model problem in function space.

3.2.1 General Dual Proximal Mappings

Similar to the update scheme for the Euclidean case from (3.1.20), we compute a full step for the Proximal Newton method at a current iterate $x \in X$ by solving the subproblem

$$\Delta x \coloneqq \operatorname*{arg\,min}_{\delta x \in X} f'(x)\delta x + \frac{1}{2} H_x(\delta x)^2 + g(x+\delta x) - g(x)\,. \tag{3.2.2}$$

If a minimizer exists, we determine the next iterate via $x_+ \coloneqq x+\Delta x$. We will consider this update scheme and investigate its convergence properties close to optimal solutions. In particular, we will be able to prove fast local convergence if H_x is adequately chosen from the image of a Newton-derivative as introduced in the context of semi-smoothness in Definition 3.1.17.

Sufficient convexity yields unique solvability of the step computation subproblem:

Proposition 3.2.1: Well-Definedness of Full Update Steps

If $\kappa_1(x) + \kappa_2 > 0$, then (3.2.2) admits a unique solution.

Proof. By assumption, the functional to be minimized is lower semi-continuous, and $\kappa_1(x) + \kappa_2 > 0$ implies that it is strictly convex as well as radially unbounded. Since X is a Hilbert space, a minimizer exists and is unique. □

Let us shortly elaborate on both constants $\kappa_1(x)$ and κ_2 as well as the assumption $\kappa_1(x)+\kappa_2 > 0$. While κ_2 is a global convexity constant for g, $\kappa_1(x)$ is a purely local quantity which differs from iterate to iterate together with the corresponding second order bilinear form H_x.

This has two immediate consequences: On the one hand, ellipticity of the second order bilinear forms can locally compensate for non-convexity of g and on the other hand (global) convexity of g enables us to locally use non-elliptic H_x even close to optimal solutions of our minimization problem. Comparing these convexity assumptions to similar works on the topic, we recognize that the authors in e.g. [44] and [55] require ellipticity of their $\nabla^2 f(x_*)$ in addition to convexity of g. In contrast, our (κ_1, κ_2)-formalism from above suitably quantifies the contribution to convexity of both f and g.

Later, for the globalization of our method, we will introduce a modification which will allow us to drop the assumption $\kappa_1(x) + \kappa_2 > 0$. For now though, we will hold on to it in order to keep the focus on local convergence behavior. For an adequate definition of proximal mappings in Hilbert space we reformulate (3.2.2) directly for the updated iterate x_+ via

$$x_+ = \operatorname*{arg\,min}_{y \in X} f'(x)(y-x) + \frac{1}{2} H_x (y-x)^2 + g(y) - g(x)\,. \tag{3.2.3}$$

In the literature, existence of a continuous inverse $H_x^{-1} : X^* \to X$ is frequently assumed, giving rise to a mapping $H_x^{-1} f' : X \to X$. Then (3.2.3) can be rearranged to

$$x_+ = \operatorname*{arg\,min}_{y \in X} g(y) + \frac{1}{2} H_x \big(y - [x - H_x^{-1} f'(x)]\big)^2\,. \tag{3.2.4}$$

In [55], this form of the updated iterate is considered and – as an adaption of (3.1.17) – the notion of a *scaled proximal mapping* $\operatorname{prox}_g^H \colon \mathbb{R}^n \to \mathbb{R}^n$ is introduced via

$$\operatorname{prox}_g^H(x) \coloneqq \operatorname*{arg\,min}_{y \in \mathbb{R}^n} g(y) + \frac{1}{2}(y-x)^T H (y-x) = \operatorname*{arg\,min}_{y \in \mathbb{R}^n} g(y) + \frac{1}{2}\big\|y - x\big\|_H^2$$

such that there (3.2.4) takes the form $x_+ = \operatorname{prox}_g^{H_x}\big(x - H_x^{-1} f'(x)\big)$. Another formal computation lets us transform first-order optimality conditions of (3.2.4) to

$$x_+ = (H_x + \partial g)^{-1} \left(H_x - f'\right) x$$

for the (in this scenario) convex subdifferential ∂g. The standing invertibility assumption on H_x then provides the equivalent formulation

$$x_+ = \left(\mathrm{Id} + H_x^{-1} \partial g\right)^{-1} \left(\mathrm{Id} - H_x^{-1} f'\right) x$$

which in particular substantiates the common interpretation of proximal-type methods as so-called *forward-backward splitting algorithms*. At first, a "forward Newton step in f" is taken followed by a "backwards subgradient step in g". Note that in particular the subdifferential of g is evaluated at the updated point x_+.

However, in this work we want to follow a different, rather direct approach towards proximal mappings which allows us to use the structure of the dual space X^* more accurately and dispense with an invertibility assumption on $H_x \in \mathcal{L}(X, X^*)$. In [104], (scaled) proximal mappings are introduced for $X = \mathbb{R}^n$ according to

$$\mathcal{P}_g^H : \mathbb{R}^n \to \mathbb{R}^n\,,\ \mathcal{P}_g^H(x) \coloneqq \operatorname*{arg\,min}_{y \in \mathbb{R}^n} g(y) + \frac{1}{2} y^T H y - x^T y\,.$$

Observing that x^T represents a dual space element in $\mathbb{R}^n$ here, we generalize this notion to the setting of Hilbert spaces and consider

$$\mathcal{P}_g^H : X^* \to X\,,\ \mathcal{P}_g^H(\varphi) \coloneqq \operatorname*{arg\,min}_{y\in X}\ g(y) + \frac{1}{2}H(y)^2 - \langle \varphi, y\rangle\,, \tag{3.2.5}$$

obtaining a mapping from the dual space back to the primal space. With this definition in mind, (3.2.3) can directly be rewritten as

$$x_+ = \operatorname*{arg\,min}_{y\in X}\ g(y) + \frac{1}{2}H_x(y)^2 - \big(H_x(x) - f'(x)\big)(y) = \mathcal{P}_g^{H_x}\big(H_x(x) - f'(x)\big)\,. \tag{3.2.6}$$

Our notion allows us to dispense with the use of the inverse H_x^{-1}, which would in addition indirectly restrict us to $\kappa_1(x) > 0$. We will refer to (3.2.5) as the *direct* or *dual formulation* of scaled proximal mappings.

We can shift convexity properties of the respective parts of the composite objective functional by inserting adequate bilinear form terms. However, this procedure does not affect the sequence of iterates generated by the update formula from above:

Lemma 3.2.2: Shifting Convexity Properties Between f and g

Let $q : X \to \mathbb{R}$ be a continuous quadratic function and denote its second derivative (which is independent of x) by $Q \coloneqq q''(x) : X \to X^*$. Consider the modified (but obviously equivalent) minimization problem

$$\begin{aligned} &\min_{x\in X} \tilde{F}(x) \coloneqq \tilde{f}(x) + \tilde{g}(x) \quad \text{where} \\ &\qquad \tilde{f}(x) \coloneqq f(x) - q(x) \quad \text{and} \quad \tilde{g}(x) \coloneqq g(x) + q(x)\,. \end{aligned} \tag{3.2.7}$$

Then, the update steps computed via (3.2.6) are identical for both problems (3.2.1) and (3.2.7) if we choose $\tilde{H}_x = H_x - Q$ as the corresponding bilinear form.

Remark. *If we choose $q(x) \coloneqq \frac{\kappa}{2}\|x\|_X^2$ for some $\kappa \in \mathbb{R}$, the modified quantities $\tilde{H}_x$ and $\tilde{g}$ suffice estimates* (A3) *and* (A4) *for $\tilde{\kappa}_1(x) = \kappa_1(x) - \kappa$ and $\tilde{\kappa}_2 = \kappa_2 + \kappa$. In particular, $\kappa_1(x) + \kappa_2 = \tilde{\kappa}_1(x) + \tilde{\kappa}_2$ remains unchanged and $\tilde{g}$ is $(\kappa + \kappa_2)$-strongly convex for $\kappa > -\kappa_2$.*

Proof. The only claim which is not apparent is the identity of update steps. To this end, we consider the fundamental definition of the update step for problem (3.2.7) at some $x \in X$ given by

$$\Delta\tilde{x} = \operatorname*{arg\,min}_{\delta x\in X} \tilde{f}'(x)\delta x + \frac{1}{2}\tilde{H}_x(\delta x)^2 + \tilde{g}(x+\delta x) - \tilde{g}(x)$$

and consequently recognize that for $q(y) = \frac{1}{2}Q(y)^2 + \ell y + c$ with constant $\ell \in X^*$ and $c \in \mathbb{R}$, the identity

$$\begin{aligned} \tilde{x}_+ &= \operatorname*{arg\,min}_{y\in X} \big(f'(x) - q'(x)\big)(y-x) + \frac{1}{2}\big(H_x - q''(x)\big)(y-x)^2 + g(y) + q(y) \\ &= \operatorname*{arg\,min}_{y\in X} \big(f'(x) - (Qx+\ell)\big)(y-x) + \frac{1}{2}\big(H_x - Q\big)(y-x)^2 + g(y) + \frac{1}{2}Q(y)^2 + \ell y \\ &= \operatorname*{arg\,min}_{y\in X} g(y) + \frac{1}{2}H_x(y)^2 - \big((H_x - Qx) - (f'(x) - Qx)\big)y \\ &= \mathcal{P}_g^{H_x}(H_x(x) - f'(x)) = x_+ \end{aligned}$$

holds, which directly shows the asserted identity of update steps. □

Remark. *If the bilinear forms for update step computation H_x and $\tilde{H}_x$ are chosen from the image of Newton-derivatives as in Definition 3.1.17 in the respective cases, we have $\tilde{H}_x = H_x - Q$, automatically.*

Remark. *In particular, only quadratic functions as above can be shifted from the smooth to the non-smooth part and vice versa without affecting update step computation. The equality of step computation subproblems crucially depends on the fact that the second order model of q above coincides with q itself. Higher order non-linearities would enter the direct decrease within the non-smooth part differently than in the respective second order model of the smooth part and thus lead to possibly different minimizers of* (3.2.2).

3.2.2 Regularity and Local Convergence Properties

The representation of the updated iterate as the image of a scaled proximal mapping in (3.2.6) will turn out to be very useful in what follows which is why we dedicate the next two propositions to the properties of scaled proximal mappings in our scenario.

Properties of Scaled Dual Proximal Mappings

The first proposition generalizes the assertions of the so-called *second prox theorem*, cf. e.g. [7, Theorem 6.39], to our notion of dual proximal mappings:

Proposition 3.2.3: General Proximal Inequality

Let H and g satisfy the assumptions (A3) and (A4) with $\kappa_1 + \kappa_2 > 0$. Then, for any $\varphi \in X^*$, the image of the corresponding proximal mapping $u := \mathcal{P}_g^H(\varphi)$ satisfies the estimate

$$\big[\varphi - H(u)\big](\xi - u) \leq g(\xi) - g(u) - \frac{\kappa_2}{2}\big\|\xi - u\big\|_X^2$$

for all $\xi \in X$.

Proof. The proof of the estimate above is an easy consequence of the characterization of the convex subdifferential of $g_H := g + \frac{1}{2}H(\cdot)^2$ together with (A4). First order conditions of the minimization problem in (3.2.5) yield

$$\varphi \in \partial\big(g + \frac{1}{2}H(\cdot)^2\big)(u) = \partial g_H(u)$$

where ∂ denotes the convex subdifferential since in particular g_H is convex due to the positivity of the sum $\kappa_1 + \kappa_2$. This inclusion directly implies the estimate

$$\varphi(y - u) + g(u) + \frac{1}{2}H(u,u) \leq g(y) + \frac{1}{2}H(y,y)$$

for arbitrary $y \in X$ which is equivalent to

$$\big[\varphi - \frac{1}{2}H(y+u)\big](y - u) \leq g(y) - g(u)\,. \tag{3.2.8}$$

As pointed out before, now we want to take advantage of the convexity assumptions on g according to (A4). To this end, we pick an arbitrary $\xi \in X$ and insert $y = y(s) := s\xi + (1-s)u$

above for $s \in]0,1]$ as an element along the straight line connecting u and ξ. Using (A4) on the right-hand side of (3.2.8) then yields

$$s\big[\varphi - H(u) - \frac{s}{2}H(\xi - u)\big](\xi - u) \leq s\big[g(\xi) - g(u) - \frac{\kappa_2}{2}(1-s)\big\|\xi - u\big\|_X^2\big]$$

where we now divide by $s \neq 0$ and subsequently evaluate the limit of s to zero. This procedure provides us with the asserted estimate for ξ, φ and u as specified above. □

The inequality from Proposition 3.2.3 can be used in order to prove several useful continuity results for general scaled proximal mappings in Hilbert spaces. However, for our purposes it suffices to assert and verify the following result, which generalizes *non-expansivity* of proximal mappings in Euclidean space, cf. [7, Theorem 6.42], to our setting. It plays a similar role as boundedness of the inverse of second order derivatives in Newton methods (, cf. Section 3.1.1,) or bounded invertibility of elements from the generalized differential within the semi-smooth Newton context (, cf. (3.1.14)).

Corollary 3.2.4: Regularity of Scaled Dual Proximal Mappings

Let H and g satisfy the assumptions (A3) and (A4) with $\kappa_1 + \kappa_2 > 0$. Then, for all $\varphi_1, \varphi_2 \in X^*$, the following Lipschitz-estimate holds:

$$\big\|\mathcal{P}_g^H(\varphi_1) - \mathcal{P}_g^H(\varphi_2)\big\|_X \leq \frac{1}{\kappa_1 + \kappa_2}\big\|\varphi_1 - \varphi_2\big\|_{X^*}.$$

Proof. Let us choose H, g, and φ_1, φ_2 as stated above. According to Proposition 3.2.3, the first order conditions for the respective minimization problems yield the inequalities

$$(\varphi_1 - H(u_1))(u_2 - u_1) \leq g(u_2) - g(u_1) - \frac{\kappa_2}{2}\big\|u_2 - u_1\big\|_X^2 \tag{3.2.9a}$$

$$(\varphi_2 - H(u_2))(u_1 - u_2) \leq g(u_1) - g(u_2) - \frac{\kappa_2}{2}\big\|u_1 - u_2\big\|_X^2 \tag{3.2.9b}$$

for $u_i = \mathcal{P}_g^H(\varphi_i)$, $i \in \{1,2\}$, since we can choose $\xi := u_2$ or $\xi := u_1$, respectively. Now, we add (3.2.9a) and (3.2.9b) which yields

$$(\varphi_2 - \varphi_1 + H(u_1 - u_2))(u_1 - u_2) \leq -\kappa_2\big\|u_1 - u_2\big\|_X^2.$$

As we rearrange this inequality, we obtain

$$H(u_1 - u_2)^2 + \kappa_2\big\|u_1 - u_2\big\|_X^2 \leq (\varphi_1 - \varphi_2)(u_1 - u_2) \leq \big\|\varphi_1 - \varphi_2\big\|_{X^*}\big\|u_1 - u_2\big\|_X$$

and eventually assumption (A3) on H yields the assertion of the proposition. □

Full Update Steps near Optimal Solutions

Even though the above continuity result for proximal mappings will turn out to be an important tool for the proof of local acceleration of the Proximal Newton method, we still have to deduce some crucial properties of the full update step Δx from (3.2.2). These will help us to characterize *optimal solutions* of (3.2.1) as fixed points of the method and then verify local acceleration afterwards. Let us shortly clarify the solution concepts which we will use in the Proximal Newton context within the following definition:

Definition 3.2.5: Stationary Points and Optimal Solutions

We call $x_* \in X$ a **stationary point** of problem (3.2.1) if the Fréchet subdifferential inclusion $0 \in \partial_F F(x_*)$ holds. By simple subdifferential calculus, this is equivalent to $-f'(x_*) \in \partial_F g(x_*)$.

By the notion of an **optimal solution** of (3.2.1), we refer to local minimizers of F on the domain of g. These solutions thus are not optimal in the sense of global minimization but in the sense of what we can achieve with the application of minimization algorithms to (3.2.1).

Let us now turn our attention back to properties of full update steps. For the first one, we generalize the notion of *descent directions* from Section 3.1.1 to the missing differentiability of our composite objective functional here. Descent directions now do not have to yield negative slope as was previously required. Instead, it is sufficient that they provide objective decrease if sufficiently scaled:

Lemma 3.2.6: Full Update Steps as Descent Directions

Suppose that f is continuously differentiable with Lipschitz derivative and that $H_x \in \mathcal{L}(X, X^*)$ suffices (A3) with $\kappa_1(x) + \kappa_2 > 0$ and κ_2 from (A4) for g. The undamped update steps computed via (3.2.2) are descent directions of the composite objective functional, i.e., the following estimate holds:

$$F(x + s\Delta x) \leq F(x) - s\big(\kappa_1(x) + \kappa_2\big)\big\|\Delta x\big\|_X^2 + O(s^2)\,.$$

Proof. Since f is assumed to be continuously differentiable and g suffices the estimate (A4), we can deduce the following bound on the composite objective functional:

$$\begin{aligned} F(x + s\Delta x) &\leq f(x) + sf'(x)\Delta x + O(s^2) \\ &\quad + sg(x + \Delta x) + (1 - s)g(x) - \frac{\kappa_2}{2}s(1 - s)\big\|\Delta x\big\|_X^2 \\ &\leq F(x) + s(f'(x)\Delta x + g(x + \Delta x) - g(x) - \frac{\kappa_2}{2}\big\|\Delta x\big\|_X^2) + O(s^2)\,. \end{aligned} \tag{3.2.10}$$

Let us now deduce an estimate for the term in brackets on the right-hand side of (3.2.10). To this end, we remember the proximal mapping representation of updated iterates in (3.2.6) and consider the corresponding estimate from Proposition 3.2.3 for $\xi := x$ which is given by

$$\big[H_x(x) - f'(x) - H_x(x_+)\big](x - x_+) \leq g(x) - g(x_+) - \frac{\kappa_2}{2}\big\|x_+ - x\big\|_X^2\,.$$

In equivalence to this statement we obtain

$$\begin{aligned} f'(x)\Delta x + g(x + \Delta x) - g(x) - \frac{\kappa_2}{2}\big\|\Delta x\big\|_X^2 &\leq -H_x(\Delta x)^2 - \kappa_2\big\|\Delta x\big\|_X^2 \\ &\leq -\big(\kappa_1(x) + \kappa_2\big)\big\|\Delta x\big\|_X^2 \end{aligned}$$

which we insert into (3.2.10) and directly obtain the asserted inequality. Note that over the course of this section we assume the positivity of the sum $\kappa_1(x) + \kappa_2$ which indeed implies from above that Δx is a descent direction. □

As mentioned beforehand, this directly enables a more insightful characterization of optimal solutions of the composite minimization problem:

Corollary 3.2.7: Optimal Solutions as Fixed Points

Consider f continuously differentiable together with $H \in \mathcal{L}(X, X^*)$ which satisfies (A3) with $\kappa_1 + \kappa_2 > 0$ and κ_2 from (A4) for g. Then, the search direction Δx_* computed according to (3.2.2) with $H_{x_*} = H$ is zero at every optimal solution $x_* \in X$ of problem (3.2.1). In particular, we obtain the fixed point equation

$$x_* = \mathcal{P}_g^H\big(H(x_*) - f'(x_*)\big)\,.$$

Proof. Since x_* is a local minimizer, we have $F(x_* + s\Delta x_*) \geq F(x_*)$ for any sufficiently small $s > 0$. By Lemma 3.2.6 this implies $\Delta x_* = 0$ for the update computed with respect to any second order bilinear form H ensuring strong convexity of subproblem (3.2.2). This then yields

$$x_* = x_* + \Delta x_* = x_{*,+} = \mathcal{P}_g^H\big(H(x_*) - f'(x_*)\big)$$

and thereby the asserted fixed point equation for x_*. □

Fast Local Convergence

Having these properties of update steps and optimal solutions in addition to the continuity result for scaled proximal mappings from Proposition 3.2.4 at hand, we can now prove the local acceleration result for our Proximal Newton method with undamped steps near optimal solutions.

For this reason, we additionally impose the following semi-smoothness assumption on $f' \colon X \to X^*$ at an optimal solution $x_* \in X$ of our problem (3.2.1): We require f' to be semi-smooth at x_* with respect to the mapping $H \colon X \to \mathcal{L}(X, X^*)$, $x \mapsto H_x$, i.e., we assume that the following approximation property holds:

$$\big\|f'(x_*) - f'(x) - H_x(x_* - x)\big\|_{X^*} = o\big(\|x - x_*\|_X\big) \quad \text{in the limit of } x \to x_*\,. \tag{3.2.11}$$

As pointed out before, this implies that adequate definitions of H_x can be taken from the image of a Newton-derivative of f' at x_* as characterized in Definition 3.1.17 for Lipschitz-continuous operators in finite dimension as well as for corresponding superposition operators, cf. Section 3.1.2.

Furthermore, let us note here that *smoothing steps* as considered in Section 3.1.2 would also fit into our semi-smooth framework here. In particular, they could serve the same purpose of providing an easier way to show semi-smoothness of the corresponding mappings also here. Since this is not a problem which we address over the course of the present treatise, we decided not to incorporate them for the sake of simplicity.

Theorem 3.2.8: Superlinear Convergence Close to Optimal Solutions

Suppose that $x_* \in X$ is an optimal solution of problem (3.2.1). Consider two consecutive iterates $x, x_+ \in X$ which have been generated by the update scheme from above and are close to x_*. Furthermore, suppose that (3.2.11) holds for H_x in addition to the standing assumptions (A1)-(A4) with $\kappa_1(x) + \kappa_2 > 0$. Then, we obtain:

$$\|x_+ - x_*\|_X = o\big(\|x - x_*\|_X\big) \quad \text{in the limit of } x \to x_*.$$

Proof. Consider the proximal mapping representations deduced above for both the updated iterate x_+ in (3.2.6) and for the optimal solution x_* in Corollary 3.2.7 with $H = H_x$ via

$$x_+ = x + \Delta x = \mathcal{P}_g^{H_x}\big(H_x(x) - f'(x)\big) \quad \text{and} \quad x_* = \mathcal{P}_g^{H_x}\big(H_x(x_*) - f'(x_*)\big)\,.$$

Next, we directly take advantage of these identities together with the continuity result for scaled proximal mappings from Proposition 3.2.4 in order to deduce the estimate

$$\begin{aligned}\big\|x_+ - x_*\big\|_X &= \big\|\mathcal{P}_g^{H_x}\big(H_x(x) - f'(x)\big) - \mathcal{P}_g^{H_x}\big(H_x(x_*) - f'(x_*)\big)\big\|_X \\ &\leq \frac{1}{\kappa_1(x) + \kappa_2}\big\|H_x(x) - f'(x) - (H_x(x_*) - f'(x_*))\big\|_{X^*} \\ &= \frac{1}{\kappa_1(x) + \kappa_2}\big\|H_x(x - x_*) - (f'(x) - f'(x_*))\big\|_{X^*} = o\big(\big\|x - x_*\big\|_X\big)\end{aligned}$$

in the limit of $x \to x_*$ where for the last identity the semi-smoothness of f' as required in (3.2.11) played a crucial role. This directly verifies the asserted local acceleration result. □

In particular, this implies local superlinear convergence of our Proximal Newton method if we can additionally verify global convergence to an optimal solution. Note that even for the local acceleration result, ellipticity of the second order approximations H_x does not necessarily have to be demanded. Also here, all that matters is strong convexity of the model of the composite functional within the subproblem (3.2.2). This might be surprising since what actually accelerates the method is the second order information on the (possibly non-convex) but differentiable part f with semi-smooth derivative f'. As a consequence, this means that the (strong) convexity of g can not only contribute to the well-definedness of update steps as solutions of (3.2.2) but also to the local acceleration of our algorithm.

The main reason for this generalization of the local acceleration result is our slightly generalized notion of proximal mappings. In particular, we have not deduced (firm) non-expansivity in the scaled norm as for example in [55] but also in that regard took advantage of the strong convexity of the composite model in the form of assumptions (A3) and (A4) with $\kappa_1(x)+\kappa_2 > 0$.

Note that for the above results to hold it has been crucial that the current iterate x is already close to an optimal solution of problem (3.2.1) which is why over the course of the next sections we will address one possibility to globalize our Proximal Newton method. We will later also realize that, eventually, we will be in the position to use undamped update steps for the computation of iterates and thereby benefit from the local acceleration result in Theorem 3.2.8.

3.2.3 Globalization via an Additional Norm Term

As we have learned over the course of the introductory section 3.1, a common way to globalize minimization algorithms is to compute fixed search directions which are then scaled by an appropriately chosen step size σ_k for the computation of the subsequent iterate. Lemma 3.2.6 from above suggests that this approach can also be taken here in case sufficient convexity of the underlying functionals is present. Such a convexity assumption, however, is one of the crucial requirements which we want to get rid of within our algorithmic formulation.

The Modified Model Decrease Functional and Damped Update Steps

To this end, let us consider an adequate augmentation of (3.2.2) and define the *damped update step* at a current iterate $x \in X$ directly as a minimizer of the *regularized second order model decrease functional* $\lambda_{x,\omega}\colon X \to]-\infty, \infty]$ which is defined by

$$\lambda_{x,\omega}(\delta x) \coloneqq f'(x)\delta x + \frac{1}{2}H_x(\delta x)^2 + \frac{\omega}{2}\|\delta x\|_X^2 + g(x+\delta x) - g(x)\,. \tag{3.2.12}$$

Following this idea, we define

$$\Delta x(\omega) \coloneqq \operatorname*{arg\,min}_{\delta x \in X} \lambda_{x,\omega}(\delta x)\,. \tag{3.2.13}$$

Here, the *regularization* (or *damping*) *parameter* $\omega \geq 0$ can be used to achieve both convexity of the second order model $\lambda_{x,\omega}$ and – as we will see later on – global convergence of the ensuing iterative method. Setting $\tilde{H} \coloneqq H_x + \omega\mathfrak{R}$ with the Riesz-Isomorphism $\mathfrak{R}\colon X \to X^*$, we observe that (A3) holds also for $\tilde{H}$ with now $\tilde{\kappa}_1(x) = \kappa_1(x) + \omega$ as the corresponding mapping. Additionally, (3.2.13) is of the form (3.2.2) which implies that the existence and regularity results of the previous sections apply using the modified quantities:

Proposition 3.2.9: Well-Definedness of Damped Update Steps

If $\omega + \kappa_1(x) + \kappa_2 > 0$ holds, then (3.2.13) admits a unique solution.

Additionally, the results of Lemma 3.2.2 apparently also hold in the globalized case. However, note that here the Hilbert space structure of X is important not only with regard to the existence of the Riesz-isomorphism but also for the strong convexity of functions of the form $g + \frac{\omega}{2}\|\cdot\|_X^2$ with g as in (A4) for arbitrary $\kappa_2 \in \mathbb{R}$. In a general Banach space setting, we can not assume additional norm terms to compensate disadvantageous convexity assumptions, cf. [7, Remark 5.18].

The updated iterate then takes the form $x_+(\omega) \coloneqq x + \Delta x(\omega)$. As a consequence of Proposition 3.2.9, for what follows, we only consider $\omega > -\big(\kappa_1(x) + \kappa_2\big)$ in order to guarantee unique solvability of the update step subproblem. The full update steps from (3.2.2) are here damped along a the so-called *proximal arc* in X which is parameterized by the regularization parameter $\omega \in]-(\kappa_1(x) + \kappa_2), \infty[$. This stands in contrast to the simple line search approach where update steps are damped along the straight line in direction of fixed $\Delta x \in X$. Apparently, the choice of the underlying norm $\|\cdot\|_X$ affects the shape of the proximal arc which additionally allows for algorithmic deliberations in that regard.

Let us now take a look at how we can rearrange the subproblem for finding an updated iterate by using the scalar product $\langle\cdot,\cdot\rangle_X$ as well as the Riesz-Isomorphism $\mathfrak{R}$:

$$\begin{aligned} x_+(\omega) &= \operatorname*{arg\,min}_{y\in X} f'(x)(y-x) + \frac{1}{2}H_x(y-x)^2 + g(y) - g(x) + \frac{\omega}{2}\|y-x\|_X^2 \\ &= \operatorname*{arg\,min}_{y\in X} g(y) + f'(x)y + \frac{1}{2}H_x(y)^2 - H_x(x,y) + \frac{\omega}{2}\|y\|_X^2 - \omega\langle x, y\rangle_X \\ &= \operatorname*{arg\,min}_{y\in X} g(y) + \frac{1}{2}\big(H_x + \omega\mathfrak{R}\big)(y)^2 - \big(H_x(x) + \omega\mathfrak{R}x - f'(x)\big)y\,. \end{aligned}$$

As a consequence of this representation, we can now summarize our update strategy in terms of scaled dual proximal mappings via

$$x_+(\omega) \coloneqq x + \Delta x(\omega) = \mathcal{P}_g^{H_x+\omega\mathfrak{R}}\big((H_x + \omega\mathfrak{R})x - f'(x)\big)\,. \tag{3.2.14}$$

We remember here that $H_x + \omega\mathfrak{R} : X \times X \to \mathbb{R}$ satisfies (A3) with constant $\kappa_1(x) + \omega$ such that the combination of g and $H_x + \omega\mathfrak{R}$ still suffices the requirements for the results from Proposition 3.2.3 and Corollary 3.2.4 for all $\omega > -(\kappa_1(x) + \kappa_2)$.

Together with the formulation of updated iterates via the above scaled proximal mapping, this enables us to establish a helpful estimate for the damped update steps $\Delta x(\omega)$:

Proposition 3.2.10: Descent Properties of Damped Update Steps

Under the assumptions (A3) for H_x and (A4) for g the inequality

$$f'(x)\Delta x(\omega) + g\big(x + \Delta x(\omega)\big) - g(x) \leq -\big(\frac{\kappa_2}{2} + \omega\big)\big\|\Delta x(\omega)\big\|_X^2 - H_x\big(\Delta x(\omega)\big)^2$$

holds for the update step $\Delta x(\omega)$ as defined in (3.2.13) and arbitrary regularization parameters $-\big(\kappa_1(x) + \kappa_2\big) < \omega < \infty$.

Proof. The proof here follows along the same lines as the derivation of the auxiliary estimate for the bracket term in the proof of Lemma 3.2.6. Due to the structure of the update formula in (3.2.14), we can take advantage of the estimate from Proposition 3.2.3 with

$$\varphi = (H_x + \omega\mathfrak{R})x - f'(x)\,, \quad H = H_x + \omega\mathfrak{R} \quad \text{and} \quad \xi = x$$

which yields $u = \mathcal{P}_g^H(x) = x_+(\omega)$ and thereby

$$H_x\big(\Delta x(\omega)\big)^2 + \omega\big\|\Delta x(\omega)\big\|_X^2 + f'(x)\Delta x(\omega) \leq g(x) - g\big(x_+(\omega)\big) - \frac{\kappa_2}{2}\big\|\Delta x(\omega)\big\|_X^2\,.$$

This inequality is equivalent to the asserted estimate. □

The Sufficient Decrease Criterion

With the above estimate for damped update steps at hand, let us now formulate a criterion for sufficient decrease which will help us to verify a global convergence result of our Proximal Newton method:

Definition 3.2.11: Admissible Regularization Parameters and Update Steps

We call a value of the regularization parameter $\omega > -\big(\kappa_1(x) + \kappa_2\big)$ and the corresponding update step $\Delta x(\omega)$ **admissible** if the **(original) sufficient decrease criterion**

$$F\big(x + \Delta x(\omega)\big) \leq F(x) + \gamma\lambda_{x,\omega}\big(\Delta x(\omega)\big) \tag{3.2.15}$$

is satisfied for some prescribed and uniform **sufficient decrease parameter** $\gamma \in]0, 1]$.

This formulation adequately generalizes the concepts of efficient step sizes from (3.1.2) and the Armijo-rule from (3.1.4) to our Proximal Newton scenario here. Furthermore, we may interpret $\lambda_\omega\big(\Delta x(\omega)\big)$ as a predicted decrease and rewrite the condition (3.2.15) as follows:

$$\frac{F\big(x + \Delta x(\omega)\big) - F(x)}{\lambda_{x,\omega}\big(\Delta x(\omega)\big)} \geq \gamma\,.$$

This is the classical ratio of actual decrease and predicted decrease which is often used for trust-region algorithms, cf. [17, Chapter 6.1]. Before now trying to verify that adequate

backtracking procedures for the above decrease criterion are again *finite*, i.e. that (3.2.15) is fulfilled for sufficiently large values of ω, we note that the assertion in Proposition 3.2.10 implies the insightful estimate

$$\begin{aligned}\lambda_{x,\omega}\big(\Delta x(\omega)\big) &\leq -\big(\frac{\kappa_2}{2}+\omega\big)\big\|\Delta x(\omega)\big\|_X^2 - \frac{1}{2}H_x\big(\Delta x(\omega)\big)^2 + \frac{\omega}{2}\|\Delta x(\omega)\|_X^2 \\ &\leq -\frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2 .\end{aligned} \tag{3.2.16}$$

By this inequality, we realize that once the sufficient decrease criterion (3.2.15) is satisfied, update steps unequal to zero provide *norm-like descent* in the composite objective functional F according to

$$F\big(x+\Delta x(\omega)\big) - F(x) \leq -\frac{\gamma}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2 . \tag{3.2.17}$$

Let us now take a look at the existence of sufficiently large values of the regularization parameter ω for (3.2.15) to be satisfied. Here, the Lipschitz-constant L_f of f' from (A1) explicitly comes into play for the first time:

Lemma 3.2.12: Satisfiability of the Sufficient Decrease Criterion (3.2.15)

For f, H_x and g as in (A1)-(A4) the criterion for sufficient decrease introduced in (3.2.15) is satisfied for $\gamma \in]0,1]$ if ω suffices

$$\omega \geq \frac{L_f - \kappa_1(x) - (1-\gamma)\big(\kappa_1(x)+\kappa_2\big)}{2-\gamma} .$$

Proof. The Lipschitz-continuity of f' directly yields the estimate

$$f(x_+(\omega)) = f(x+\Delta x(\omega)) \leq f(x) + f'(x)\Delta x(\omega) + \frac{L_f}{2}\big\|\Delta x(\omega)\big\|_X^2$$

from where we immediately obtain an estimate for the descent in the composite objective functional via

$$\begin{aligned}F\big(x_+(\omega)\big) - F(x) &\leq f'(x)\Delta x(\omega) + \frac{L_f}{2}\big\|\Delta x(\omega)\big\|_X^2 + g(x_+(\omega)) - g(x) \\ &= f'(x)\Delta x(\omega) + \frac{1}{2}(H_x+\omega\mathfrak{R})\big(\Delta x(\omega)\big) + g(x_+(\omega)) - g(x) \\ &\quad + \frac{L_f-\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 - \frac{1}{2}H_x\big(\Delta x(\omega)\big)^2 \\ &\leq \lambda_\omega(\Delta x(\omega)) + \frac{L_f-\kappa_1(x)-\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 .\end{aligned} \tag{3.2.18}$$

By our lower bound on ω from the assertion and by (3.2.16) we obtain

$$\frac{L_f-\kappa_1(x)-\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 \leq \frac{1-\gamma}{2}(\omega+\kappa_1(x)+\kappa_2)\big\|\Delta x(\omega)\big\|_X^2 \leq -(1-\gamma)\lambda_{x,\omega}\big(\Delta x(\omega)\big) .$$

Inserted above this yields

$$F\big(x_+(\omega)\big) - F(x) \leq \lambda_{x,\omega}\big(\Delta x(\omega)\big) - (1-\gamma)\lambda_{x,\omega}\big(\Delta x(\omega)\big) = \gamma\lambda_{x,\omega}\big(\Delta x(\omega)\big) .$$

This estimate is equivalent to (3.2.15) and thereby concludes the proof of the assertion. □

Remark. *Let us shortly remark on a special case, in which a different bound for ω than formulated above suffices for admissibility. In the last step of* (3.2.18) *the bound*

$$\omega \geq L_f - \kappa_1(x) \tag{3.2.19}$$

allows us to omit the norm term and use the non-positivity of $\lambda_{x,\omega}\big(\Delta x(\omega)\big)$ according to (3.2.16) *in order to achieve the sufficient decrease criterion. The bounds coincide for $\gamma = 1$ and, in general,* (3.2.19) *relaxes the one from Lemma 3.2.12 if*

$$\frac{L_f - \kappa_1(x) - (1-\gamma)\big(\kappa_1(x) + \kappa_2\big)}{2-\gamma} > L_f - \kappa_1(x)$$

holds which is equivalent to $\kappa_2 < -L_f$. In particular, note that κ_2 is generally perceived as the largest constant such that (A4) *holds. This means that every smaller $\tilde{\kappa}_2$ also suffices the requirement and can be used for any of the estimates which we deduce from* (A4)*. Mostly, however, κ_2 only appears in order to estimate images of dual proximal mappings with respect to g. In those estimates positivity of $\omega + \kappa_1(x) + \kappa_2$ has always been crucial for well-definedness of occurring quantities. Thus, artificially considering $\tilde{\kappa_2} < \kappa_2$ also yields an implicit increase of ω which makes the ensuing bound on ω facile. In order to obtain a meaningful bound on the actually used regularization parameter, we should always refer to κ_2 as the largest constant for which* (A4) *is satisfied.*

On this note, let us also remark here that all of the bounds on ω deduced above are invariant under shifting convexity as formulated in Lemma 3.2.2 since the Lipschitz constant L_f of f' also changes when adding/subtracting quadratic norm terms to the respective parts of the objective functional.

Global Convergence by Regularization

In particular, the above result implies the boundedness of the sequence of regularization parameters (ω_k) as long as these are increased by no more than a factor in case the sufficient decrease criterion (3.2.15) fails. Additionally, for global convergence, similar as we did for admissible step sizes in Definition 3.1.2 in the framework of Gradient methods, we have to guarantee that

$$\lambda_{x,\omega_k}\big(\Delta x(\omega_k)\big) \to 0 \quad \text{implies} \quad \big\|\Delta x(\omega_k)\big\|_X \to 0 \tag{3.2.20}$$

in order to achieve global convergence. A simple way to guarantee this is to prescribe an upper bound $\overline{M} > 0$ and with that impose the restriction

$$\big\|\Delta x(\omega_k)\big\|_X^2 \leq -\overline{M}\lambda_{x_k,\omega_k}\big(\Delta x(\omega_k)\big)\,. \tag{3.2.21}$$

Let us shortly discuss the relation of this additional requirement for update steps with the similar bound (3.2.16) from before. By the assumption of well-definedness of our update steps via $\omega + \kappa_1(x) + \kappa_2 > 0$, it is tempting to think that (3.2.21) is unnecessary. In that regard, consider the scenario where $\kappa_1(x) + \kappa_2 < 0$ holds close to a stationary point of (3.2.1). Then, the prefactor in (3.2.16) might deteriorate which destroys the implication from (3.2.20). Thus, we have to additionally impose a *uniformity condition* for that estimate as done in (3.2.21). On the other hand, (3.2.16) is still helpful in the context by providing (3.2.21) if ω is chosen sufficiently large. Furthermore, in case we have $(\kappa_1(x) + \kappa_2) > 1/\overline{M}$ (, e.g. close to optimal

solutions), also $\omega = 0$ suffices in that regard. Generally, $\overline{M}$ constitutes an algorithmic safety parameter which we can choose very large in practice. All in all, this results in the following algorithm:

Algorithm 9: Second Order Semi-Smooth Proximal Newton Method

Data: Starting point $x_0 \in \operatorname{dom} g$, sufficient decrease parameter $\gamma \in]0, 1]$, initial value $\omega_0 \geq 0$, threshold $\varepsilon > 0$ for stopping criterion
Initialization: $k = 0$;
while $(1 + \omega_k)\left\|\Delta x_k(\omega_k)\right\|_X \geq \varepsilon$ **do**
  Compute a trial step $\Delta x_k(\omega_k)$ according to (3.2.13);
  while *bound* (3.2.21) *or sufficient decrease criterion* (3.2.15) *is not satisfied* **do**
    Increase regularization parameter ω_k adequately;
    Recompute trial step $\Delta x_k(\omega_k)$ as above;
  end
  Update the current iterate to $x_{k+1} \leftarrow x_k + \Delta x_k(\omega_k)$;
  Decrease ω_k appropriately to some $\omega_{k+1} < \omega_k$ for next iteration;
  Update the sequence index $k \leftarrow k + 1$;
end

Now that we have formulated the algorithm and can be sure that we can always damp update steps sufficiently such that they yield decrease according to (3.2.15), our next goal is to verify the stationarity of limit points of the sequence of iterates generated by Algorithm 9. To this end, we will first prove that the norm of the corresponding update steps converges to zero along the sequence of iterates if F is bounded from below:

Lemma 3.2.13: Convergence of Update Step Norms

Let $(x_k) \subset X$ be the sequence generated by the Proximal Newton method from Algorithm 9. Then, either $F(x_k) \to -\infty$ or $\left(\lambda_{x_k,\omega_k}\left(\Delta x_k(\omega_k)\right)\right)_{k\in\mathbb{N}}$ and thereby $\left(\left\|\Delta x_k(\omega_k)\right\|_X\right)_{k\in\mathbb{N}}$ converge to zero for $k \to \infty$.

Proof. By (3.2.17), the sequence $F(x_k)$ is monotonically decreasing. Thus, either $F(x_k) \to -\infty$ or $F(x_k) \to \underline{F}$ for some $\underline{F} \in \mathbb{R}$ and thus in particular $F(x_k) - F(x_{k+1}) \to 0$. Since $\gamma > 0$ holds, we also have $\lambda_{x_k,\omega_k}(\Delta x(\omega_k)) \to 0$. From there, assumption (3.2.21) immediately implies $\left\|\Delta x_k(\omega_k)\right\|_X \to 0$ as discussed beforehand. □

In the following, we will assume throughout that $\left(F(x_k)\right)_{k\in\mathbb{N}}$ is bounded from below. As a consequence, the result from Lemma 3.2.13 enables us to take further steps towards global convergence. The distance of $0 \in X^*$ to the Fréchet-subdifferential of F can be considered a meaningful quantity in the context. We formulate its limit behavior in the following proposition.

Proposition 3.2.14: Convergence of the Fréchet Subdifferential Distance

Along the sequence generated by the Proximal Newton method from Algorithm 9 we have $\operatorname{dist}(\partial_F F(x_k), 0) \to 0$ for $k \to \infty$.

Proof. Taking a look at the optimality conditions for update step computation in (3.2.13) at some $x \in X$, we recognize

$$(H_x + \omega\mathfrak{R})x - f'(x) \in \partial g_\omega^{H_x}(x_+(\omega))$$

with the convex subdifferential of $g_\omega^{H_x} : X \to]-\infty, \infty], y \mapsto g(y) + \frac{1}{2}H_x(y)^2 + \frac{\omega}{2}\|y\|_X^2$ on the right-hand side. This directly yields the existence of some Fréchet-subderivative $\eta \in \partial_F g(x_+(\omega))$ such that

$$\eta + f'(x_+(\omega)) = r_{x,\omega}\big(\Delta x(\omega)\big) \quad \text{with} \quad r_{x,\omega}(v) \coloneqq f'(x+v) - f'(x) - \big(H_x + \omega\mathfrak{R}\big)v\,. \tag{3.2.22}$$

For any sequence index $k \in \mathbb{N}$, this implies the estimate

$$\operatorname{dist}(\partial_F F(x_{k+1}), 0) = \operatorname{dist}\big(f'(x_{k+1}) + \partial_F g(x_{k+1}), 0\big) \leq \big\|r_{x_k,\omega_k}\big(\Delta x_k(\omega_k)\big)\big\|_{X^*}\,.$$

Thus, by Lemma 3.2.13 and

$$\big\|r_{x_k.\omega_k}\big(\Delta x_k(\omega_k)\big)\big\|_{X^*} \leq \big(L_f + \big\|H_{x_k}\big\|_{\mathcal{L}(X,X^*)} + \omega_k\big)\big\|\Delta x_k(\omega_k)\big\|_X$$

together with the bounds on bilinear form norms from (A2) and on regularization parameters from the remark following Lemma 3.2.12 we obtain the desired convergence. □

The above estimates let us conclude that we can indeed interpret $\big\|\Delta x_k(\omega_k)\big\|_X \leq \varepsilon$ as a condition for the optimality of the subsequent iterate up to some prescribed accuracy. However, small step norms $\big\|\Delta x_k(\omega_k)\big\|_X$ can also occur due to very large values of the damping parameter ω_k, cf. Lemma 4.2.1. As a consequence, the algorithm would stop even though the sequence of iterates has not yet reached an optimal solution of the problem. In order to rule out this inconvenient case, we consider the scaled version $(1+\omega_k)\big\|\Delta x_k(\omega_k)\big\|_X$ as the stopping criterion in Algorithm 9.

Now, we are in the position to discuss at least subsequential convergence of our algorithm to a stationary point as introduced in Definition 3.2.5. We start with the case of convergence in norm:

Theorem 3.2.15: Stationarity of Accumulation Points

Under the standing assumptions (A1)-(A4), all accumulation points $\bar{x} \in X$ (in norm) of the sequence of iterates (x_k) generated by the Proximal Newton method from Algorithm 9 are stationary points of problem (3.2.1).

Proof. For the sake of notational simplicity, we will by $(x_k)_{k\in\mathbb{N}}$ refer to the subsequence which converges to the accumulation point $\bar{x}$ in norm.

Let us consider a modified version of our minimization problem as in (3.2.7) in Lemma 3.2.2 and choose $q(x) = \frac{1}{2}Q(x)^2$ for $Q : X \times X \to \mathbb{R}$ such that $\tilde{g} = g + q$ is (strongly) convex on its domain. This is always possible by (A4). According to Lemma 3.2.2, the sequence of iterates remains unchanged and step computation takes the form

$$x_{k+1} = \tilde{x}_{k+1} = \operatorname*{arg\,min}_{y\in X} \tilde{g}(y) + \frac{1}{2}\big(H_{x_k} - Q + \omega\mathfrak{R}\big)(y)^2 - \big((H_{x_k} + \omega_k\mathfrak{R})x_k - f'(x_k)\big)y$$

with first order optimality conditions

$$(H_{x_k} + \omega_k \mathfrak{R})x_k - f'(x_k) \in \partial\tilde{g}(x_{k+1}) + \big(H_{x_k} - Q + \omega\mathfrak{R}\big)(x_{k+1})$$

where $\partial\tilde{g}(x_{k+1})$ denotes the convex subdifferential of $\tilde{g}$ at x_{k+1}. Consequently, we know that there exists some $\tilde{\eta}_k \in \partial\tilde{g}(x_{k+1})$ such that

$$\tilde{\eta}_k + \big(f'(x_{k+1}) - Qx_{k+1}\big) = r_{x_k}\big(\Delta x_k(\omega_k)\big)$$

holds with the remainder term on the right-hand side as before given by

$$r_{x,\omega}(v) \coloneqq f'(x+v) - f'(x) - \big(H_x + \omega\mathfrak{R}\big)v\,.$$

Again, this remainder term $r_{x_k.\omega_k}\big(\Delta x_k(\omega_k)\big) \in X^*$ tends to zero for $k \to \infty$. In particular, this allows us to conclude the existence of the limit expression $\tilde{\eta} \coloneqq \lim_{k\to\infty} \tilde{\eta}_k = -f'(\bar{x}) + Q\bar{x}$ by the continuity of f' and the convergence of (x_k) in norm. The definition of the convex subdifferential $\partial\tilde{g}$ together with the lower semi-continuity of $\tilde{g}$ directly yields

$$\begin{aligned}
\tilde{g}(\xi) - \tilde{g}(\bar{x}) &= \tilde{g}(\xi) - g(\bar{x}) - \frac{1}{2}Q(\bar{x})^2 \geq \tilde{g}(\xi) - \liminf_{k\to\infty} g(x_k) - \lim_{k\to\infty} \frac{1}{2}Q(x_k)^2 \\
&= \liminf_{k\to\infty} \tilde{g}(\xi) - \tilde{g}(x_k) \geq \liminf_{k\to\infty} \tilde{\eta}_k(\xi - x_k) \\
&= \lim_{k\to\infty} \tilde{\eta}_k(\xi - x_k) = \tilde{\eta}(\xi - \bar{x})
\end{aligned}$$

for any $\xi \in X$ which proves the inclusion $\tilde{\eta} \in \partial\tilde{g}(\bar{x})$. The evaluation of the latter limit expression can easily be retraced by splitting

$$\tilde{\eta}_k(\xi - x_k) = \tilde{\eta}_k(\xi - \bar{x}) + (\tilde{\eta}_k - \tilde{\eta})(\bar{x} - x_k) + \tilde{\eta}(\bar{x} - x_k)\,. \tag{3.2.23}$$

In particular, we recognize $\tilde{\eta} \in \partial\tilde{g}(\bar{x})$ as $-f'(\bar{x}) + Q\bar{x} \in \partial\tilde{g}(\bar{x})$ and equivalently $-f'(\bar{x}) \in \partial_F g(\bar{x})$ for the Frechét-subdifferential ∂_F. This implies $0 \in \partial_F F(\bar{x})$, i.e., the stationarity of our limit point $\bar{x}$ as in Definition 3.2.5. □

Also note that – in general – the above global convergence result does not rely on the strong convexity of the composite objective function F but yields stationarity of limit points also in the non-convex case of $\kappa_1(x_k) + \kappa_2 < 0$ if $\omega_k > -\big(\kappa_1(x_k) + \kappa_2\big)$ is chosen adequately. In particular, this ensures that also independent of strong convexity assumptions near optimal solutions, the algorithm approaches the optimal solution and can then benefit from additional convexity at later iterations.

While bounded sequences in finite dimensional spaces always have convergent subsequences, we can only expect *weak subsequential convergence* in general Hilbert spaces. As one consequence, existence of minimizers of non-convex functions on Hilbert spaces can usually only be established in the presence of some compactness. On this count, we note that in (3.2.23) even weak convergence of $x_k \rightharpoonup \bar{x}$ would be sufficient. Unfortunately, in the latter case we can not evaluate the limit of $f'(x_k) \to f'(\bar{x})$.

In order to extend our proof to this situation, we require some more structure for both of the parts of our composite objective functional. To this end, we remember the following well-known definition of compact operators:

Definition 3.2.16: Compactness of Linear Operators

A linear operator $K : X \to Y$ between two normed vector spaces X and Y is called **compact** if one of the following equivalent statements holds:

(i) The image of the unit ball of X is relatively compact in Y (, i.e., its closure is compact).

(ii) For any bounded sequence $(x_n)_{n\in\mathbb{N}} \subset X$ the image sequence $(Kx_n)_{n\in\mathbb{N}} \subset Y$ contains a strongly convergent subsequence $\left(x_{n_k}\right)_{k\in\mathbb{N}} \subset X$.

With this notion at hand, we can formulate the following global convergence theorem:

Theorem 3.2.17: Global Convergence Under Additional Structural Assumptions

Let f be of the form $f(x) = \hat{f}(x)+\check{f}(Kx)$ where K is a compact operator. Additionally, assume that $g + \hat{f}$ is convex and weakly lower semi-continuous in a neighborhood of stationary points of (3.2.1). Then, weak convergence of the sequence of iterates $x_k \rightharpoonup \bar{x}$ suffices for $\bar{x} \in X$ to be a stationary point of (3.2.1).

If F is strictly convex and radially unbounded, the whole sequence x_k converges weakly to the unique minimizer $x_* \in X$ of F. If F is κ-strongly convex, with $\kappa > 0$, then $x_k \to x_*$ in norm.

Proof. We can employ the same proof as above replacing g by $g+\hat{f}$ and using that due to the additional compactness assumption, we now have $\tilde{f}'(Kx_k) \to \check{f}'(K\bar{x})$ in norm. Finally, this then shows

$$(g + \hat{f})(u) - (g + \hat{f})(\bar{x}) \geq \eta(u - \bar{x}),$$

i.e., $\eta = -\check{f}'(K\bar{x})K \in \partial(g + \hat{f})(\bar{x}) = \partial_F g(\bar{x}) + \{\hat{f}'(\bar{x})\}$ which in particular implies

$$-f'(\bar{x}) = -\check{f}'(K\bar{x})K - \hat{f}'(\bar{x}) \in \partial_F g(\bar{x}).$$

This again constitutes $0 \in \partial_F F(\bar{x})$ and thereby the stationarity of the weak limit point $\bar{x}$.

Let us now consider the second assertion: F being strictly convex as well as radially unbounded yields that problem (3.2.1) has a unique solution x_*. Additionally, we know that our sequence of iterates is bounded as a consequence of which we can select a weakly convergent subsequence. The first assertion of the theorem then implies that the limit of each subsequence we choose is a stationary point of problem (3.2.1), and thus by convexity the unique optimal solution x_*. A standard weak subsubsequence argument then shows that the whole sequence converges to x_* weakly.

If F is κ-strongly convex, then – as discussed below (A4) – the diameter of the level sets $L_{F(x_k)}$ tends to 0 as $k \to \infty$, since $F(x_k) \to F(x_*)$. This implies $\|x_k - x_*\|_X \to 0$. □

Remark. *Note that here we did not shift the "non-compact" part of the smooth part f to g in the sense of Lemma 3.2.2 but considered the (apparently equivalent) composite optimization problem with smooth part $\check{f}(K\cdot)$ and non-smooth part $g + \hat{f}$. Even though the ensuing Proximal Newton method generally leads to a different sequence of iterates than for the original formulation, global convergence follows by the proof from above.*

As it is to be expected, additional structural assumptions on the composite objective functional like convexity and compactness lead to stronger convergence results concerning the sequence of iterates generated by our method. However, the stationarity result from Theorem 3.2.15, which has been formulated under minimal structural assumptions, is satisfying in the context and enables application of Algorithm 9 to a large variety of problems.

3.2.4 Second Order Semi-Smoothness

In order to be able to benefit from the local acceleration result in Theorem 3.2.8, we have to ensure that under our standing assumptions on F, eventually also full steps are admissible for sufficient decrease according to the criterion formulated in (3.2.15). To this end, we will introduce a new notion of differentiability, which we call *second order semi-smoothness*, and investigate how it interacts with our Proximal Newton method.

For the smooth part f of our composite objective function F we define a second order semi-smoothness estimate at some $x \in \mathrm{dom} f$ by the *second order approximation property*

$$f(x+\xi) = f(x) + f'(x)\xi + \frac{1}{2}H_{x+\xi}(\xi,\xi) + o(\|\xi\|_X^2) \quad \text{for } \xi \to 0\,. \tag{3.2.24}$$

This will be precisely the assumption which we need in order to conclude the *transition to fast local convergence* in the following section.

General Definition and Calculus

We will now give a general definition for operators and develop the standard results to be expected in the context: Denote by $\mathcal{L}^{(2)}(X,Y)$ the normed space of bounded vector valued bilinear forms $X \times X \to Y$, equipped with the usual norm defined by

$$\|B\|_{\mathcal{L}^{(2)}(X,Y)} := \sup_{\xi_1,\xi_2 \neq 0} \frac{\|B(\xi_1,\xi_2)\|_Y}{\|\xi_1\|_X \|\xi_2\|_X}\,.$$

Our definition intuitively lifts the approximation property, which determines semi-smoothness in Definition 3.1.16, to the second order level. This also motivates the name of the concept:

Definition 3.2.18: Second Order Semi-Smoothness of Continuously Differentiable Operators

Let X, Y be Banach Spaces and let $V \subset X$ be an open subset on which the operator $T\colon V \to Y$ is defined. Consider a set-valued mapping $\partial^{(2)}T\colon V \rightrightarrows \mathcal{L}^{(2)}(X,Y)$ with non-empty images, i.e., $\partial^{(2)}T(x) \neq \emptyset$ for all $x \in V$.

(a) We say that T is $\partial^{(2)}T$-**second-order-semi-smooth** at $x \in V$ if T is continuously differentiable near x and the following approximation property holds:

$$\sup_{\mathcal{M} \in \partial^{(2)}T(x+\xi)} \|T(x+\xi) - T(x) - T'(x)\xi - \frac{1}{2}\mathcal{M}(\xi,\xi)\|_Y = o(\|\xi\|_X^2) \quad \text{for } \xi \to 0\,.$$

(b) Accordingly, we refer to $\partial^{(2)}T\colon V \rightrightarrows \mathcal{L}^{(2)}(X,Y)$ as the **generalized second order differential** of T. We will always assume non-emptiness of its images. Implicitely, $\partial^{(2)}T$-second-order-semi-smoothness of T at $x \in V$ shall automatically imply non-emptiness of images at least in a neighborhood of x.

As was the case for the generalized differential in first order semi-smoothness from Definition 3.1.16, the specific choice of $\partial^{(2)}T$ is far from unique and depends on the application at hand. As pointed out beforehand, our algorithmic approach to the concept for transition to local convergence of Proximal Newton methods benefits more from a rather direct characterization via image elements of the generalized second order differential. Again here, we avoid the peculiarities arising from defining a general notion of Newton-differentiability and directly generalize the concept of Newton-derivatives from Definition 3.1.17 to our second order scenario:

Definition 3.2.19: Second Order Semi-Smoothness with Respect to an Operator

Consider Banach spaces X, Y, an open subset $V \subset X$, a point $x \in V$, and a neighborhood $N(x) \subset V$ of x. The continuously differentiable mapping $T\colon V \to Y$ is called **second order semi-smooth** at x **with respect to the operator** $\mathcal{T}\colon N(x) \to \mathcal{L}^{(2)}(X,Y)$ if the following approximation property holds:

$$\big\|T(x+\xi) - T(x) - T'(x)\xi - \frac{1}{2}\mathcal{T}(x+\xi)(\xi,\xi)\big\|_Y = o\big(\|\xi\|_X^2\big) \quad \text{for } \xi \to 0\,.$$

We then call $\mathcal{T}$ a **second order Newton-derivative** of F at $x \in V$.

As was the case in the first order scenario, this notion is intimately connected to the one of $\partial^{(2)}T$-second-order-semi-smoothness from Definition 3.2.18: The generalized second order differential of a continuously differentiable and $\partial^{(2)}T$-second-order-semi-smooth mapping F as before allows us to define mappings with respect to which F is second order semi-smooth. Having such a mapping in place, on the other hand, enables the construction of a generalized second order differential $\partial^{(2)}T$ such that F is then $\partial^{(2)}T$-second-order-semi-smooth.

In what follows, we will develop a standard set of calculus rules for our notion of second order semi-smoothness with respect to operators. For a start, twice continuously differentiable operators apparently are second order semi-smooth:

Proposition 3.2.20: Second Order Semi-Smoothness by Classical Second Order Differentiability

Assume that $T\colon V \to Y$ is twice continuously differentiable at $x \in V$. Then, T is second order semi-smooth at x with respect to the ordinary second order derivative T''.

Proof. While continuous differentiability of T is apparent, the second order approximation property follows by a simple computation:

$$\begin{aligned} &T(x-\xi) - \big[T(x) + T'(x)\xi + \frac{1}{2}T''(x+\xi)(\xi,\xi)\big] \\ &\qquad = \big[T(x-\xi) - T(x) - T'(x)\xi - \frac{1}{2}T''(x)(\xi,\xi)\big] + \frac{1}{2}\big[T''(x) - T''(x+\xi)\big](\xi,\xi)\,. \end{aligned}$$

In their respective Y-norm, both terms in square brackets are $o\big(\|\xi\|_X^2\big)$. The first by Fréchet differentiability of T, the second by continuity of T''. □

It is an obvious remark that the sum of two second order semi-smooth functions is second order semi-smooth again with linear and quadratic terms defined via sums. Furthermore, the following chain rule can be shown:

Theorem 3.2.21: Second Order Semi-smooth Chain Rule

Consider Banach spaces X, Y and Z together with open subsets $D_S \subset X$, $D_T \subset Y$. Additionally, suppose that $S\colon D_S \to Y$ and $T\colon D_T \to Z$ with $S(D_S) \subset D_T$ are second order semi-smooth at $x \in D_S$ and $y = S(x) \in D_T$ with respect to $\mathcal{S}$ and $\mathcal{T}$, respectively. Then $T \circ S$ is second order semi-smooth with respect to $\mathcal{TS}$ defined as follows:

$$\mathcal{TS}(x)(\xi_1, \xi_2) \coloneqq \mathcal{T}(y)(S'(x)\xi_1, S'(x)\xi_2) + T'(y)\mathcal{S}(x)(\xi_1, \xi_2)\,.$$

Proof. Apparently, $T \circ S$ is continuously differentiable by the classical chain rule. For the proof of the non-trivial approximation property, we introduce the notations $x_\xi = x + \xi$, $y_\xi = S(x_\xi)$, and $\eta = y_\xi - y$. With these prerequisites we can, as usual for chain rules, split the remainder term as follows:

$$\begin{aligned}
&(T \circ S)(x_\xi) - (T \circ S)(x) - (T \circ S)'(x)\xi - \frac{1}{2}\mathcal{TS}(x_\xi)(\xi, \xi) \\
&\quad = T(y_\xi) - T(y) - T'(y)S'(x)\xi \\
&\qquad - \frac{1}{2}\Big(\mathcal{T}(y_\xi)(S'(x_\xi)\xi, S'(x_\xi)\xi) + T'(y_\xi)\mathcal{S}(x_\xi)(\xi, \xi)\Big) \\
&\quad = T(y_\xi) - T(y) - T'(y)\eta - \frac{1}{2}\mathcal{T}(y_\xi)(\eta, \eta) && (3.2.25) \\
&\qquad + T'(y)\left(S(x_\xi) - S(x) - S'(x)\xi - \frac{1}{2}\mathcal{S}(x_\xi)(\xi, \xi)\right) && (3.2.26) \\
&\qquad + \frac{1}{2}(T'(y) - T'(y_\xi))\mathcal{S}(x_\xi)(\xi, \xi) && (3.2.27) \\
&\qquad + \frac{1}{2}\big(\mathcal{T}(y_\xi)(\eta, \eta) - \mathcal{T}(y_\xi)\big(S'(x_\xi)\xi, S'(x_\xi)\xi\big)\big)\,. && (3.2.28)
\end{aligned}$$

We will show that each of the expressions (3.2.25)-(3.2.28) is $o\big(\|\xi\|_X^2\big)$: For (3.2.25) this follows from second order semi-smoothness of T while second order semi-smoothness of S implies the desired result for (3.2.26). Continuity of T' and boundedness of $\mathcal{S}$ yield that (3.2.27) is $o\big(\|\xi\|_X^2\big)$. Finally, (3.2.28) can be reformulated via the third binomial formula:

$$\begin{aligned}
\big\|\mathcal{T}(y_\xi)(\eta, \eta) - \mathcal{T}(y_\xi)\big(S'(x_\xi)\xi, S'(x_\xi)\xi\big)\big\|_Z &= \big\|\mathcal{T}(y_\xi)\big(\eta + S'(x_\xi)\xi, \eta - S'(x_\xi)\xi\big)\big\|_Z \\
&\leq \big\|\mathcal{T}(y_\xi)\big\|_{\mathcal{L}^{(2)}(Y,Z)}\big\|\eta + S'(x_\xi)\xi\big\|_Y\big\|\eta - S'(x_\xi)\xi\big\|_Y\,.
\end{aligned}$$

By continuous differentiablity of S (which is a prerequisite of second order semi-smoothness by our definition), we estimate $\big\|\eta + S'(x_\xi)\xi\big\|_Y = O\big(\|\xi\|_X\big)$ together with

$$\big\|\eta - S'(x_\xi)\xi\big\|_Y \leq \big\|\eta - S'(x)\xi\big\|_Y + \big\|(S'(x) - S'(x_\xi))\xi\big\|_Y = o\big(\|\xi\|_X\big)$$

which finally yields the desired result. □

Remark. *In the case $T'(y) = 0$, we observe from* (3.2.26) *that S only needs to be continuously differentiable and we may set $\mathcal{S} = 0$.*

Second order semi-smoothness of T and semi-smoothness of T' as in (3.2.11) are closely related but not equivalent in general. In case we additionally have continuity of the Newton-derivative, we have the following connection if we choose $\mathcal{T}$ from Definition 3.2.19 as a Newton-derivative of T' in the sense of Definition 3.1.17:

Proposition 3.2.22: Second Order Semi-Smoothness by Semi-Smoothness

Let the operator $T\colon V \to Y$, $V \subset X$ open, be continuously differentiable and let $T'\colon V \to \mathcal{L}(X,Y)$ be semi-smooth at $x \in V$ with respect to the Newton-derivative $\mathcal{T}\colon X \to \mathcal{L}^{(2)}(X,Y)$. Furthermore, assume that $\mathcal{T}$ is continuous at x.

Then, T is second order semi-smooth at x with respect to $\mathcal{T}$ as a second order Newton-derivative.

Proof. By the fundamental theorem of calculus and the continuous differentiability of T, we obtain the following identity for any $\xi \in X$ such that $x+\xi \in V$:

$$\begin{aligned}
&T(x+\xi) - T(x) - T'(x)\xi - \frac{1}{2}\mathcal{T}(x+\xi)(\xi,\xi) \\
&\quad= \int_0^1 T'(x+s\xi)\xi \,\mathrm{d}s - T'(x)\xi - \frac{1}{2}\mathcal{T}(x+\xi)(\xi,\xi) \\
&\quad= \int_0^1 \big[T'(x+s\xi) - T'(x) - \mathcal{T}(x+\xi)(s\xi)\big]\xi \,\mathrm{d}s \\
&\quad= \int_0^1 \big[T'(x+s\xi) - T'(x) - \mathcal{T}(x+s\xi)(s\xi)\big]\xi \,\mathrm{d}s + \int_0^1 \big[\mathcal{T}(x+s\xi) - \mathcal{T}(x+\xi)\big](s\xi,\xi)\,\mathrm{d}s\,.
\end{aligned}$$

In norm, it is easy to see with the mean value theorem for integration that the latter two expressions are both $o(\|\xi\|_X^2)$ in the limit of $\xi \to 0$: the first one by semi-smoothness of T' with respect to $\mathcal{T}$ and the second one by continuity of $\mathcal{T}$. □

Remark. *Conversely, we cannot deduce semi-smoothness from second order semi-smoothness even under the additional continuity assumption for the second order Newton-derivative. All we can achieve in that scenario is an estimate of the form*

$$\big\|T'(x+\xi)\xi - T'(x)\xi - H_{x+\xi}(\xi,\xi)\big\|_Y = o\big(\|\xi\|_X^2\big)$$

in the limit of $\xi \to 0$ which in particular does not imply the approximation property for semi-smoothness in sufficient generality.

Without the continuity of the Newton-derivative, however, this connection between second order semi-smoothness of T and semi-smoothness of T' can not be established. Let us shortly give a both simple and illustrative example: Consider the function

$$h\colon \mathbb{R} \to \mathbb{R}\,, \quad h(x) \coloneqq \begin{cases} x^3 \sin\left(\frac{1}{x}\right) & ,\ x \neq 0 \\ 0 & ,\ x = 0 \end{cases}$$

which is continuously differentiable with $h'(x) = x\big[3x\sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right)\big]$, $x \neq 0$, and $h'(0) = 0$. The cubic asymptotics of h suggest that $\mathcal{H} \equiv 0$ is a possible definition for second order semi-smoothness of h at $x_* = 0$ as above. Apparently, we obtain for $x \in \mathbb{R}$ and $\delta x = x - x_* = x$:

$$|h(x) - h(x_*) - h'(x_*)\delta x - \frac{1}{2}\mathcal{H}(x)(\delta x)^2| = |\delta x|^3\big|\sin\left(\frac{1}{x}\right)\big| = O\big(|\delta x|^3\big) \ \text{ for } \ \delta x \to 0\,,$$

i.e., that h is indeed second order semi-smooth at $x_* = 0$ with respect to $\mathcal{H}$. On the other hand, we have

$$|h'(x_*) - h'(x) - \mathcal{H}(x)(x_* - x)| = |\delta x|\big|3x\sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right)\big| \neq o\big(|\delta x|\big) \ \text{ for } \ \delta x \to 0\,,$$

which implies that h' is indeed not semi-smooth at $x_* = 0$ with respect to the same $\mathcal{H}$. However, in many cases of practical interest, both conditions can be shown to hold.

Second Order Semi-smoothness of Superposition Operators

The function $\phi\colon \mathbb{R} \to \mathbb{R},\ \phi(x) \coloneqq \max\{0, x\}^2$, for instance, is of particular importance in the optimization literature, e.g. for the reformulation of certain optimal control problems with the help of the introduction of adjoint states, cf. [97, Section 6]. These reformulations are an interesting alternative which then naturally describe algorithms based on the discretization idea of [42]. The real valued variant of ϕ here is second order semi-smooth at the point $x = 0$ with respect to

$$\mathcal{P}\colon \mathbb{R} \to \mathbb{R}\,, \quad \mathcal{P}(x) \coloneqq \begin{cases} 0 & ,\ x < 0 \\ 1 & ,\ x \geq 0 \end{cases} \tag{3.2.29}$$

as well as twice Fréchet differentiable (and thus also second-order semi-smooth, cf. Proposition 3.2.20) at any other point $x \neq 0$ with the same $\phi'' = \mathcal{P}$.[17] In its most common use in optimization with PDE's, however, ϕ is mostly used in the form of a superposition operator the concept of which we have shortly elaborated on towards the end of Section 3.1.2.

For this reason, it seems fitting to investigate how we can lift second order semi-smoothness as introduced in Definition 3.2.19 to superposition operators on L^p-spaces for appropriate p. Fortunately, it turns out that this is possible and that we are able to transfer standard arguments in the context to our second order scenario.

For convenience, we recapitulate the following lemma, which is a slight generalization of a standard result on the continuity of superposition operators. For further context and a comprehensible proof, we refer to [97, Lemma 3.1].

Lemma 3.2.23: Continuity of Superposition Operators

Let Ω be a measurable subset of $\mathbb{R}^d$ and $\psi\colon \mathbb{R} \times \Omega \to \mathbb{R}$. For each measurable function $x\colon \Omega \to \mathbb{R}$, assume that the mapping $\Psi(x)$ defined by $\Psi(x)(\omega) = \psi(x(\omega), \omega)$ is measurable. Let $x \in L^p(\Omega, \mathbb{R})$ be given. Then, the following assertion holds:

If ψ is continuous in the first component at $(x(\omega), \omega)$ for almost all $\omega \in \Omega$, and Ψ maps $L^p(\Omega, \mathbb{R})$ into $L^s(\Omega, \mathbb{R})$ for $1 \leq p, s < \infty$, then Ψ is continuous at x in the norm topology.

The standard text book result, cf. [117, Proposition 26.7(a)] or [25, Proposition IV.1.1], requires ψ to be a Carathéodory function as introduced prior to (2.2.23), and thus in particular continuous in the first component for all $t \in \Omega$. This assumption is slightly weakened here to the almost everywhere sense. It is known, for example, that pointwise limits and suprema of Carathéodory functions yield superposition operators that map measurable functions to measurable functions, cf. e.g. [95, p. 21.4]. This class of mappings is also referred to as the one of *Baire-Carathéodory* functions. The mapping $\mathcal{P}$ from (3.2.29) is an example.

As has been done in its original appearance in [97], the above continuity result for superposition operators can be used in several scenarios to show convergence of accordingly defined remainder term operators. Thus, it constitutes a both easily comprehensive and powerful tool

[17] Here and in the following, we identify both bilinear forms from $\mathcal{L}^{(2)}(\mathbb{R}, \mathbb{R})$ and linear mappings from $\mathcal{L}(\mathbb{R}, \mathbb{R})$ with the corresponding real number in $\mathbb{R}$.

to prove all sorts of smoothness results of arbitrarily involved superposition operators. This approach adequately simplifies the treatment of such problems.

Since our ensuing results considering second order semi-smoothness of superposition operators crucially depend on the assumptions made for the exponents s and p in Lemma 3.2.23, it seems reasonable to shortly discuss them here as was done in [97, Remark 3.2]:

(i) In order to ensure that Ψ maps $L^p(\Omega)$ into $L^s(\Omega)$ one has to verify the following growth condition for ψ:

$$|\psi(x,\omega)| \leq a(\omega) + b|x|^{p/s} \quad \text{for some } a \in L^s(\Omega) \text{ and } b \in \mathbb{R}\,.$$

(ii) In case we have $p < \infty$ and $\big\|\Psi(x)\big\|_{L^\infty(\Omega)} \leq M$ uniformly for all $x \in L^p(\Omega)$, the proof of Lemma 3.2.23 simplifies but its assertions still only hold for all $s < \infty$ and *not* for $s = \infty$ (except for the case of constant Ψ). In particular, this will result in a so-called *norm gap* which is commonly observed in the analysis of semi-smooth Newton methods, cf. [111, Example 3.57]. For $s = \infty$, only a weak form of the above continuity result can be shown, cf. [97, Section 5].

(iii) For $p = \infty$ and $s < \infty$, the proof of Lemma 3.2.23 carries over in a modified way. The case $p = s = \infty$, on the other hand, can not be handled analogously. Here, the continuity of ψ at $x(t)$ in the first component has to be uniform in Ω which is usually a too strong assumption in the context of semi-smoothness.

With this continuity result at hand, we can now investigate how second order semi-smoothness of real valued mappings transfers to the correspondingly defined superposition operators. It turns out that also here known concepts from first order semi-smoothness theory shine through:

Proposition 3.2.24: Second Order Semi-smoothness of Superposition Operators

Consider a real-valued function $\phi\colon \mathbb{R} \to \mathbb{R}$ with globally Lipschitz-continuous derivative $\phi'\colon \mathbb{R} \to \mathbb{R}$. Suppose that ϕ is second order semi-smooth with respect to the bounded mapping $\mathcal{P}\colon \mathbb{R} \to \mathbb{R}$.
Let $\Omega \subset \mathbb{R}^d$ be a set of finite measure and assume that the composition $\mathcal{P} \circ u\colon \Omega \to \mathbb{R}$ is measurable for any measurable function $u : \Omega \to \mathbb{R}$.

Then, for each $p > 2$ and $x \in L^p(\Omega)$, the superposition operator $\Phi : L^p(\Omega) \to L^1(\Omega)$, $\Phi(x)(\omega) \coloneqq \phi\big(x(\omega)\big)$, is second order semi-smooth at x with respect to the operator

$$\mathfrak{P}\colon L^p(\Omega) \to \mathcal{L}^{(2)}(L^p(\Omega), L^1(\Omega))\,, \quad \mathfrak{P}(x)(\xi_1,\xi_2)(\omega) = \mathcal{P}(x(\omega))\xi_1(\omega)\xi_2(\omega)\,.$$

Proof. Let us all across the proof here consider a representative $x \in L^p(\Omega)$ with $p > 2$. For the first time in this section, continuous differentiability of the superposition operator Φ at x is not apparent which is why we use the opportunity to showcase the proof of such a result: Intuitively, we define our candidate for the Fréchet-derivative $\Phi'\colon L^p(\Omega) \to \mathcal{L}\big(L^p(\Omega), L^1(\Omega)\big)$ by $\Phi'(x)(\omega) \coloneqq \phi'\big(x(\omega)\big)$ for any $\omega \in \Omega$. In order to now show that

$$\big\|\Phi(x+\xi) - \Phi(x) - \Phi'(x)\xi\big\|_{L^1(\Omega)} = o\big(\big\|\xi\big\|_{L^p(\Omega)}\big) \quad \text{for } \xi \to 0 \text{ in } L^p(\Omega) \tag{3.2.30}$$

holds at x from above, we define the remainder term function $\tilde{r}_x\colon \mathbb{R}\times\Omega\to\mathbb{R}$ by

$$\tilde{r}_x(t,\omega) \coloneqq \frac{\phi\big(x(\omega)+t\big)-\phi\big(x(\omega)\big)-\phi'\big(x(\omega)\big)t}{t}$$

for $t\neq 0$ completed by $\tilde{r}_x(0,\omega)\coloneqq 0$. By assumption, ϕ is continuously differentiable with globally Lipschitz derivative ϕ' with constant L_ϕ. This provides us with $|\tilde{r}_x(t,\omega)|\leq\frac{L_\phi}{2}t$ which directly implies continuity of $\tilde{r}_x$ in the first component at $t=0$ for almost all $\omega\in\Omega$. In particular, the bound on $|\tilde{r}_x|$ from above and the finiteness of Ω in measure allows for the definition of the corresponding superposition operator

$$\tilde{R}_x\colon L^p(\Omega)\to L^s(\Omega)\,,\quad \tilde{R}_x(\xi)(\omega)\coloneqq\tilde{r}_x\big(\xi(\omega),\omega\big)$$

at least for any $1\leq s\leq p$. By Lemma 3.2.23, we infer the continuity of $\tilde{R}_x$ at $\xi=0\in L^p(\Omega)$ at least for finite such s. We can use this desired property of the remainder term superposition operator insofar that Hölder's inequality with $1/p+1/s=1$ now[18] yields

$$\big\|\Phi(x+\xi)-\Phi(x)-\Phi'(x)\xi\big\|_{L^1(\Omega)}=\big\|\tilde{R}_x(\xi)\cdot\xi\big\|_{L^1(\Omega)}\leq\big\|\tilde{R}_x(\xi)\big\|_{L^s(\Omega)}\big\|\xi\big\|_{L^p(\Omega)}=o\big(\big\|\xi\big\|_{L^p(\Omega)}\big)$$

in the limit of $\xi\to 0$ in $L^p(\Omega)$ and thus the desired identity (3.2.30). Continuity of the derivative can be handled in the same way by considering

$$\hat{r}_x\colon\mathbb{R}\times\Omega\to\mathbb{R}\,,\quad \hat{r}_x(t,\omega)\coloneqq\phi'\big(x(\omega)+t\big)-\phi'\big(x(\omega)\big)$$

together with its superposition operator and again Lemma 3.2.23.

With continuous differentiability out of the way, we can take a similar route in order to prove the approximation property characterizing second order semi-smoothness in Definition 3.2.19. This time around, we consider the remainder term function $r_x\colon\mathbb{R}\times\Omega\to\mathbb{R}$ defined by

$$r_x(t,\omega)\coloneqq\frac{\phi(x(\omega)+t)-\phi(x(\omega))-\phi'(x(\omega))t-\mathcal{P}(x(\omega)+t)t^2}{t^2}$$

for $t\neq 0$ and $r_x(t,\omega)\coloneqq 0$ for $t=0$. By Lipschitz-continuity of ϕ' with constant L_ϕ and boundedness $\mathcal{P}<C$ for some $C>0$ we observe that r_x is bounded uniformly on $\mathbb{R}\times\Omega$:

$$|r_x(t,\omega)|\leq\frac{1}{t^2}\big(\frac{L_\phi}{2}t^2+Ct^2\big)=\frac{L_\phi}{2}+C<\infty\,.$$

Together with the finiteness of Ω in measure, this lets us conclude that the superposition operator $R_x\colon L^p(\Omega)\to L^s(\Omega)$, $R_x(\xi)(\omega)=r_x(\xi(\omega),\omega)$ is well-defined for any $1\leq s\leq\infty$. Second order semi-smoothness of ϕ with respect to $\mathcal{P}$ then yields

$$\big|\phi(x(\omega)+t)-\phi(x(\omega))-\phi'(x(\omega))t-\mathcal{P}(x(\omega)+t)t^2\big|=o(t^2)\quad\text{in the limit of } t\to 0$$

from which we infer continuity of $r_x(\cdot,\omega)$ at $t=0$ for almost all $\omega\in\Omega$. Hence, by Lemma 3.2.23, R_x is continuous as an operator at $\xi=0\in L^p(\Omega)$ for any $s<\infty$. By Hölder's inequality with $1/s+2/p=1$, we deduce that

$$\begin{aligned}\big\|\Phi(x+\xi)-\Phi(x)-\Phi'(x)\xi-\mathfrak{P}(x+\xi)(\xi,\xi)\big\|_{L^1(\Omega)}&=\big\|R_x(\xi)\cdot\xi\cdot\xi\big\|_{L^1(\Omega)}\\&\leq\big\|R_x(\xi)\big\|_{L^s(\Omega)}\big\|\xi\big\|^2_{L^p(\Omega)}=o\big(\big\|\xi\big\|^2_{L^p(\Omega)}\big)\end{aligned}$$

holds in the limit of $\xi\to 0$ in $L^p(\Omega)$. Establishing the above estimate concludes the proof of second order semi-smoothness of Φ at x. □

[18] In particular, together with $p>2$ this implies $s<p$ which is crucial for the well-definedness of the superposition operator.

As we have seen over the course of the proof of the above assertions, the continuity result from Lemma 3.2.23 provides a flexible framework for the consideration of all kinds of smoothness properties of superposition operators. In particular, the continuity of remainder terms can directly be tied to said differentiability statements. As a consequence, the assumptions on the Lebesgue exponent p in Proposition 3.2.24 immediately follow the ones imposed on p and s in Lemma 3.2.23. Since we have continuity of the remainder term only for $s < \infty$, Hölder's inequality restricts p to be strictly larger than two. Unsurprisingly and in analogy to the theory of semi-smooth superposition operators, this results in a norm gap in the sense that Proposition 3.2.24 is false for $p = 2$. This is closely related to the so called *two-norm discrepancy* which illustrates the sometimes peculiar relationship of remainder terms and accordingly chosen norms in order to show crucial properties of functionals in concrete applications (, cf. e.g. [106, Section 4.10.2]).

In the above example of the squared max-function, $\mathcal{P}$ from (3.2.29) has a discontinuity at $x = 0$, so also in general we can not expect that the corresponding superposition operator $\mathfrak{P}$ is a continuous mapping on a given open set. However, we can show the following result:

Proposition 3.2.25: Continuity of Second Order Approximations of Superposition Operators

Let $p > 2$ and a representative $x \in L^p(\Omega)$ be fixed. Consider some $\mathcal{P}\colon \mathbb{R} \to \mathbb{R}$ and assume that the function $\mathcal{P}_x\colon \mathbb{R}\times\Omega \to \mathbb{R}$ given by $\mathcal{P}_x(t,\omega) \coloneqq \mathcal{P}\big(x(\omega)+t\big)$ is continuous in $t = 0$ for almost all $\omega \in \Omega$.

Then, the mapping $\mathfrak{P}\colon L^p(\Omega) \to \mathcal{L}^{(2)}(L^p(\Omega), L^1(\Omega))$, for $\omega \in \Omega$ accordingly defined via $\mathfrak{P}(x)(\xi_1,\xi_2)(\omega) = \mathcal{P}(x(\omega))\xi_1(\omega)\xi_2(\omega)$, is continuous at x.

Proof. By the continuity assumptions on $\mathcal{P}$, we can directly apply Lemma 3.2.23 to the superposition operator $\tilde{\mathfrak{P}}\colon L^p(\Omega) \to L^s(\Omega)$, $\tilde{\mathfrak{P}}(x)(\omega) \coloneqq \mathcal{P}(x(\omega))$, for $s < \infty$. Again, Hölder's inequality with $1/s + 2/p = 1$ helps us out in order to obtain

$$\begin{aligned}\big\|\mathfrak{P}(\tilde{x}) - \mathfrak{P}(x)\big\|_{\mathcal{L}^{(2)}(L^p(\Omega),L^1(\Omega))} &= \sup_{\xi_1,\xi_2\neq 0} \frac{\big\|\big(\mathfrak{P}(\tilde{x}) - \mathfrak{P}(x)\big)(\xi_1,\xi_2)\big\|_{L^1(\Omega)}}{\|\xi_1\|_{L^p(\Omega)}\|\xi_2\|_{L^p(\Omega)}} \\ &\leq \big\|\tilde{\mathfrak{P}}(\tilde{x}) - \tilde{\mathfrak{P}}(x)\big\|_{L^s(\Omega)}\end{aligned}$$

for any $x, \tilde{x} \in L^p(\Omega)$ which allows us to conclude the continuity of $\mathfrak{P}$ from the one of $\tilde{\mathfrak{P}}$. □

In our example $\phi(x) = \max\{0, x\}^2$, the corresponding second order Newton derivative $\mathcal{P}$ from (3.2.29) fulfills the hypothesis of this theorem at $x \in L^p(\Omega)$, if $x = 0$ only on a set of measure zero in Ω. This kind of regularity assumption can also be found frequently in the literature on semi-smooth Newton methods. For an example, consider [40], where the authors reformulate a Mixed Complementarity Problem (MCP) for the application of semi-smooth Newton methods and proof corresponding mesh-independence results. In the context of non-smooth equations obtained from reformulations of MCPs, the non-differentiability points are those where strict complementarity is violated. For the corresponding analysis, it is sufficient that the set where strict complementarity is violated is a set of measure zero.

3.2.5 Transition to Fast Local Convergence

Let us now turn our attention back to our Proximal Newton method. In order to benefit from the local acceleration result from Theorem 3.2.8, we have to consider the admissibility of undamped update steps with respect to the sufficient decrease criterion as formulated in (3.2.15) near optimal solutions of problem (3.2.1). This concept first appeared in our introductory description of Newton methods in Lemma 3.1.9. In our non-smooth setting here, both the semi-smoothness of f' from (3.2.11) and the second order semi-smoothness of f from (3.2.24) will contribute a crucial part to the proof of this result.

However, an algorithm that tests in every iterate whether the undamped Proximal Newton step is acceptable will most likely compute many unnecessary trial iterates, in particular during the early phase of globalization. Thus, it additionally is of interest whether damped Proximal Newton steps are acceptable as well close to the solution.

In order to establish the corresponding proposition of admissibility, we will first have to more closely investigate the influence of the regularization parameter ω on the norm of steps computed according to (3.2.13). The following monotonicity results gives valuable insight into this dependence:

Lemma 3.2.26: Monotonicity of Update Step Norms Regarding Regularization Parameters

Let $\Delta x(\omega)$ and $\Delta x(\tilde{\omega})$ be exactly computed update steps at an iterate $x \in X$ according to (3.2.13) with regularization parameters satisfying $\omega > -(\kappa_1(x) + \kappa_2)$ and $\tilde{\omega} \geq \omega$.

Then, the following norm estimates hold:

$$\big\|\Delta x(\omega) - \Delta x(\tilde{\omega})\big\|_X \leq \frac{\tilde{\omega} - \omega}{\omega + \kappa_1(x) + \kappa_2} \big\|\Delta x(\tilde{\omega})\big\|_X \,, \tag{3.2.31}$$

$$\big\|\Delta x(\tilde{\omega})\big\|_X \leq \big\|\Delta x(\omega)\big\|_X \leq \frac{\tilde{\omega} + \kappa_1(x) + \kappa_2}{\omega + \kappa_1(x) + \kappa_2} \big\|\Delta x(\tilde{\omega})\big\|_X \,. \tag{3.2.32}$$

Proof. We consider the proximal representation of exactly computed update steps

$$x + \Delta x(\hat{\omega}) = x_+(\hat{\omega}) = \mathcal{P}_g^{H_x + \hat{\omega}\mathfrak{R}}\big((H_x + \hat{\omega}\mathfrak{R})x - f'(x)\big)$$

for $\hat{\omega} \in \{\omega, \tilde{\omega}\}$. Via Proposition 3.2.3, from these we can deduce the respective proximal inequalities

$$\begin{aligned}\big[\big(\hat{\omega}\mathfrak{R} + H_x\big)x - f'(x) - \big(\hat{\omega}\mathfrak{R} + H_x\big)x_+(\hat{\omega})\big]\big(\hat{\xi} - x_+(\hat{\omega})\big) \\ \leq g(\hat{\xi}) - g\big(x_+(\hat{\omega})\big) - \frac{\kappa_2}{2}\big\|\hat{\xi} - x_+(\hat{\omega})\big\|_X^2\end{aligned} \tag{3.2.33}$$

for any $\hat{\xi} \in X$ which we choose as $\hat{\xi} = x_+(\hat{\omega})$ for the respectively other $\hat{\omega} \in \{\omega, \tilde{\omega}\}$ and add the ensuing estimates in order to obtain

$$\big[\big(\omega\mathfrak{R} + H_x\big)\Delta x(\omega) - \big(\tilde{\omega}\mathfrak{R} + H_x\big)\Delta x(\tilde{\omega})\big]\big(\Delta x(\omega) - \Delta x(\tilde{\omega})\big) \leq -\kappa_2\big\|\Delta x(\omega) - \Delta x(\tilde{\omega})\big\|_X^2 \,.$$

We now insert a $(\omega\mathfrak{R} + H_x)\Delta x(\tilde{\omega})$-term to the left-hand squared bracket and simplify which yields

$$\begin{aligned}\big(\omega\mathfrak{R} + H_x\big)\big(\Delta x(\omega) - \Delta x(\tilde{\omega})\big)^2 + \kappa_2\big\|\Delta x(\omega) - \Delta x(\tilde{\omega})\big\|_X^2 \\ \leq (\tilde{\omega} - \omega)\mathfrak{R}\big(\Delta x(\tilde{\omega}), \Delta x(\omega) - \Delta x(\tilde{\omega})\big)\end{aligned}$$

where we can now additionally utilize (A3) for the simpler form

$$(\omega + \kappa_1(x) + \kappa_2)\big\|\Delta x(\omega) - \Delta x(\tilde{\omega})\big\|_X^2 \leq (\tilde{\omega} - \omega)\Re\big(\Delta x(\tilde{\omega}), \Delta x(\omega) - \Delta x(\tilde{\omega})\big)\,. \tag{3.2.34}$$

From here, we can take two paths both of which contribute to the completion of the proof. Firstly, we divide by $(\omega + \kappa_1(x) + \kappa_2) > 0$ and use the Cauchy-Schwarz-Inequality on the right-hand side which then implies

$$\big\|\Delta x(\omega) - \Delta x(\tilde{\omega})\big\|_X^2 \leq \frac{\tilde{\omega} - \omega}{\omega + \kappa_1(x) + \kappa_2}\big\|\Delta x(\tilde{\omega})\big\|_X\big\|\Delta x(\omega) - \Delta x(\tilde{\omega})\big\|_X\,,$$

i.e., exactly (3.2.31) since the difference norm term can be assumed to be non-zero without loss of generality. Moving on, we take advantage of

$$\big\|\Delta x(\omega)\big\|_X \leq \big\|\Delta x(\omega) - \Delta x(\tilde{\omega})\big\|_X + \big\|\Delta x(\tilde{\omega})\big\|_X \leq \big(1 + \frac{\tilde{\omega} - \omega}{\omega + \kappa_1(x) + \kappa_2}\big)\big\|\Delta x(\tilde{\omega})\big\|_X$$

and thereby directly obtain the second inequality from (3.2.32).

The other way to manipulate (3.2.34) is to simply drop the left-hand side due to $(\omega + \kappa_1 + \kappa_2) > 0$. This immediately yields

$$(\tilde{\omega} - \omega)\big\|\Delta x(\tilde{\omega})\big\|_X^2 \leq (\tilde{\omega} - \omega)\Re\big(\Delta x(\tilde{\omega}), \Delta x(\omega)\big)$$

where we use the Cauchy-Schwarz-Inequality and divide by $(\tilde{\omega} - \omega)\big\|\Delta x(\tilde{\omega})\big\|_X$ which again can be assumed to be non-zero (and positive) without loss of generality. The ensuing estimate then constitutes the first part of (3.2.32), completing the proof. □

In particular, this monotonicity result of the update step norms with respect to the regularization parameter also incorporates the case of $\omega = 0$ in the strongly convex scenario of $\kappa_1(x) + \kappa_2 > 0$ which we assumed for local acceleration close to optimal solutions of (3.2.1). This also enables us to prove the following corollary concerning the limit behavior of damped update steps:

Corollary 3.2.27: Limit Behavior of Damped Exact Update Steps

For $x \in X$ close to an optimal solution x_* of (3.2.1) with $\kappa_1(x) + \kappa_2 > 0$, we can find constants $c_1, c_2 > 0$ such that for $\omega \geq 0$ and $x_+(\omega) = x + \Delta x(\omega)$, the following estimates hold:

$$\big\|x_+(\omega) - x_*\big\|_X \leq c_1\big\|x - x_*\big\|_X \quad \text{and} \quad \big\|x - x_*\big\|_X \leq c_2\big\|\Delta x(\omega)\big\|_X\,.$$

Remark. *In particular, these eventual norm estimates have implications on the limit behavior of the respective terms. If we now have $\xi = o\big(\|x_+(\omega) - x_*\|_X\big)$ for some $\xi \in X$, $\xi = o\big(\|x - x_*\|_X\big)$ immediately holds and from there we obtain $\xi = o\big(\|\Delta x(\omega)\|_X\big)$ analogously in the limit of the current iterate approaching the solution.*

Proof. For the deduction of both asserted inequalities, we will take advantage of the local superlinear convergence stated in Theorem 3.2.8, i.e., $\big\|x_+ - x_*\big\|_X = o\big(\big\|x - x_*\big\|_X\big)$ in the limit of $x \to x_*$. Consequently, we can write

$$\big\|x_+ - x_*\big\|_X = \psi\big(\big\|x - x_*\big\|_X\big)\big\|x - x_*\big\|_X \tag{3.2.35}$$

for some modulus of continuity $\psi : [0,\infty[\to [0,\infty[$ with $\psi(t) \to 0$ for $t \to 0$. With this helpful representation at hand, we estimate

$$\begin{aligned}\|x_+(\omega) - x_*\|_X &\leq \|x - x_*\|_X + \|\Delta x(\omega)\|_X \leq \|x - x_*\|_X + \|\Delta x\|_X \\ &\leq 2\|x - x_*\|_X + \|x_+ - x_*\|_X = \big[2 + \psi\big(\|x - x_*\|_X\big)\big]\|x - x_*\|_X \,.\end{aligned}$$

By the definition of ψ above, this directly implies the first asserted inequality. We can deduce the second one similarly quickly via

$$\|x - x_*\|_X \leq \|x_+ - x_*\|_X + \|\Delta x\|_X = \psi\big(\|x - x_*\|_X\big)\|x - x_*\|_X + \|\Delta x\|_X \,.$$

We can assume $\psi\big(\|x - x_*\|_X\big) < 1$ close to the optimal solution x_* and thereby deduce

$$\begin{aligned}\|x - x_*\|_X &\leq \big[1 - \psi\big(\|x - x_*\|_X\big)\big]^{-1}\|\Delta x\|_X \\ &\leq \big[1 - \psi\big(\|x - x_*\|_X\big)\big]^{-1}\Big(\frac{\omega}{\kappa_1 + \kappa_2} + 1\Big)\|\Delta x(\omega)\|_X\end{aligned}$$

with the additional help of (3.2.32). Taking into account that ω remains bounded completes the proof of the second asserted inequality. □

Now, we are in the position to prove the admissibility of both undamped and damped steps close to optimal solutions of the composite minimization problem (3.2.1). We will see that undamped steps will generally be admissible whereas for the admissibility of damped steps we will have to assume the following additional property of the second order model bilinear forms:

$$(H_{x_+(\omega)} - H_x)(x_+(\omega) - x_*)^2 = o\big(\|x - x_*\|_X^2\big) \quad \text{in the limit of } x \to x_* \,. \tag{3.2.36}$$

In the following proposition, we will both recognize the necessity of this estimate and give sufficient conditions under which it is satisfied:

Proposition 3.2.28: Admissibility of Exact Update Steps Close to Optimal Solutions

Let $x_* \in X$ be an optimal solution of (3.2.1) and let H_x suffice (A3) as well as g suffice (A4) with $\kappa_1(x) + \kappa_2 > 0$ in a neighborhood of x_*. Additionally, suppose that (3.2.24) holds for f as well as (3.2.11) holds for f' at x_* with respect to the mapping $H\colon X \to \mathcal{L}^{(2)}(X,\mathbb{R})$, $x \mapsto H_x$, which satisfies (3.2.36).

Then, for any $\gamma \in]0,1]$ and $\omega \geq 0$, we can find some neighborhood $U_{\gamma,\omega} \subset X$ of x_* such that at all $x \in U_{\gamma,\omega}$ the corresponding update step $\Delta x(\omega)$ from (3.2.13) is admissible for sufficient decrease according to (3.2.15) for that γ.

In particular:

(i) Full steps Δx as defined in (3.2.2) are eventually admissible.

(ii) If the (second order) Newton-derivative $H\colon X \to \mathcal{L}^{(2)}(X,\mathbb{R})$, $x \mapsto H_x$, is continuous at $x = x_*$, then eventually all steps are admissible.

Remark. *More precisely, we have to demand that $f' \colon X \to X^*$ is semi-smooth with respect to $H \colon X \to \mathcal{L}(X, X^*)$ and not with respect to $H \colon X \to \mathcal{L}^{(2)}(X, \mathbb{R})$. For the sake of notational simplicity, however, we will identify these mappings when stating assumptions here and in what follows.*

Proof. Let us take a look at the descent in the composite objective function F when performing an update step and see which estimates we can deduce with the help of the assumptions and results preceding this proposition.

We will denote the update by $\Delta x(\omega)$ or $x_+(\omega) = x + \Delta x(\omega)$ respectively for some arbitrary $\omega \geq 0$ such that the notation comprises both the damped and undamped case for the update step. Now, we simply expand the difference in F via

$$F\big(x + \Delta x(\omega)\big) - F(x) = f\big(x + \Delta x(\omega)\big) - f(x) + g\big(x + \Delta x(\omega)\big) - g(x)$$

and estimate the descent in the smooth part of the objective function $f\big(x + \Delta x(\omega)\big) - f(x)$. By telescoping we obtain the identity

$$\begin{aligned}
&f\big(x_+(\omega)\big) - f(x) - f'(x)\Delta x(\omega) - \frac{1}{2} H_x\big(\Delta x(\omega)\big)^2 \\
&\quad = f\big(x_+(\omega)\big) - f(x_*) - f'(x_*)\big(x_+(\omega) - x_*\big) - \frac{1}{2} H_{x_+(\omega)}\big(x_+(\omega) - x_*\big)^2 \\
&\qquad + f(x_*) + f'(x_*)(x - x_*) + \frac{1}{2}(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2 \\
&\qquad - f(x) - H_x\big(\Delta x(\omega)\big)^2 + \frac{1}{2}\big[H_x\big(x_* - x_+(\omega)\big)^2 + H_x\big(\Delta x(\omega)\big)^2\big] \\
&\qquad - f'(x)\Delta x(\omega) + f'(x_*)\Delta x(\omega) + H_x\big(x - x_*, \Delta x(\omega)\big) \\
&\qquad + H_x\big(x_* - x_+(\omega) + \Delta x(\omega), \Delta x(\omega)\big)
\end{aligned}$$

which we can then reformulate in a more revealing way via

$$\begin{aligned}
&f\big(x_+(\omega)\big) - f(x) - f'(x)\Delta x(\omega) - \frac{1}{2} H_x\big(\Delta x(\omega)\big)^2 \\
&\quad = \left[f\big(x_+(\omega)\big) - f(x_*) - f'(x_*)\big(x_+(\omega) - x_*\big) - \frac{1}{2} H_{x_+(\omega)}\big(x_+(\omega) - x_*\big)^2\right] \\
&\qquad - \left[f(x) - f(x_*) - f'(x_*)(x - x_*) - \frac{1}{2} H_x(x - x_*)^2\right] \\
&\qquad - \Big[\big(f'(x) - f'(x_*)\big) - H_x(x - x_*)\Big]\Delta x(\omega) + \frac{1}{2}(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2 \\
&\quad = o\big(\|x_+(\omega) - x_*\|_X^2\big) + o\big(\|x - x_*\|_X^2\big) + o\big(\|x - x_*\|_X\big)\|\Delta x(\omega)\|_X \\
&\qquad + \frac{1}{2}(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2 .
\end{aligned} \tag{3.2.37}$$

In the last step, we have used second order semi-smoothness of f from (3.2.24) and semi-smoothness of f' at x_* from (3.2.11) – both with respect to $H \colon X \to \mathcal{L}^{(2)}(X, \mathbb{R})$.

We observe that the only critical term is $\rho(x, \omega) \coloneqq \frac{1}{2}(H_{x_+(\omega)} - H_x)(x_+(\omega) - x_*)^2$. We conclude

$$f\big(x + \Delta x(\omega)\big) - f(x) = f'(x)\Delta x(\omega) + \frac{1}{2} H_x\big(\Delta x(\omega)\big)^2 + \rho(x, \omega) + o\big(\|\Delta x(\omega)\|_X^2\big)$$

by Corollary 3.2.27 and from there directly deduce

$$F\big(x_+(\omega)\big) - F(x) = \lambda_{x,\omega}\big(\Delta x(\omega)\big) - \frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 + \rho(x,\omega) + o\big(\big\|\Delta x(\omega)\big\|_X^2\big)\,.$$

This estimate allows us to now study the relation between the actual descent in F and the second order model $\lambda_{x,\omega}$.

We choose $\gamma \in]0,1]$ and $\omega \geq 0$ and look for the neighborhood $U_{\gamma,\omega}$ such that $\Delta x(\omega)$ computed at any $x \in U_{\gamma,\omega}$ yields sufficient decrease according to (3.2.15) for that γ. To this end, we define the *decrease ratio function* $\gamma\colon X \times [0,\infty[\to]-\infty,\infty]$ by the fraction

$$\begin{aligned}\gamma(x,\omega) \coloneqq \frac{F\big(x_+(\omega)\big) - F(x)}{\lambda_{x,\omega}\big(\Delta x(\omega)\big)} &= 1 + \frac{-\frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 + \rho(x,\omega) + o\big(\big\|\Delta x(\omega)\big\|_X^2\big)}{\lambda_{x,\omega}\big(\Delta x(\omega)\big)} \\ &= 1 + \frac{\frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 - o\big(\big\|\Delta x(\omega)\big\|_X^2\big) - \rho(x,\omega)}{\big|\lambda_{x,\omega}\big(\Delta x(\omega)\big)\big|}\end{aligned}$$

which (eventually) should be larger than the previously chosen and fixed γ.

We may assume that the numerator of the latter expression is non-positive, otherwise this inequality is trivially fulfilled. Thus, by decreasing the positive denominator via (3.2.16), we obtain

$$\gamma(x,\omega) \geq 1 + \frac{\omega}{\omega + \kappa_1(x) + \kappa_2} - \frac{\rho(x,\omega) + o\big(\big\|\Delta x(\omega)\big\|_X^2\big)}{\frac{1}{2}\big(\omega + \kappa_1(x) + \kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2}\,. \tag{3.2.38}$$

Since the limit $x \to x_*$ also yields $\Delta x(\omega) \to 0$ which can again be retraced via

$$\big\|\Delta x(\omega)\big\|_X \leq \big\|\Delta x\big\|_X \leq \big\|x_+ - x_*\big\|_X + \big\|x - x_*\big\|_X$$

and due to the limit assumption (3.2.36) for the $\rho(x,\omega)$-term, we conclude that we can choose the neighborhood $U_{\gamma,\omega}$ as the one where the latter fraction term in (3.2.38) suffices

$$\frac{o\big(\big\|\Delta x(\omega)\big\|_X^2\big) + \rho(x,\omega)}{\big\|\Delta x(\omega)\big\|_X^2} < \frac{1}{2}\big[\big(\omega + \kappa_1(x) + \kappa_2\big)(1-\gamma) + \omega\big]\,.$$

Inserting this bound into the estimate (3.2.38), we immediately obtain that $\Delta x(\omega)$ computed at any $x \in U_{\gamma,\omega}$ is admissible for sufficient decrease according to (3.2.15) for γ chosen beforehand.

The $\rho(x,\omega)$-term vanishing by assumption (3.2.36) is in particular implied by either (i) or (ii) in the following way:

$$\begin{aligned}\text{(i)} \quad \Rightarrow \quad &\big|(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2\big| = \big|(H_{x_+} - H_x)(x_+ - x_*)^2\big| \\ &\leq \big(\big\|H_{x_+}\big\|_{\mathcal{L}^{(2)}(X,\mathbb{R})} + \big\|H_x\big\|_{\mathcal{L}^{(2)}(X,\mathbb{R})}\big)\big\|x_+ - x_*\big\|_X^2 = o\big(\big\|x - x_*\big\|_X^2\big)\,, \\ \text{(ii)} \quad \Rightarrow \quad &\big|(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2\big| \\ &\leq \big(\big\|H_{x_+(\omega)} - H_{x_*}\big\|_{\mathcal{L}^{(2)}(X,\mathbb{R})} + \big\|H_{x_*} - H_x\big\|_{\mathcal{L}^{(2)}(X,\mathbb{R})}\big)\big\|x_+(\omega) - x_*\big\|_X^2 \\ &= o\big(\big\|x - x_*\big\|_X^2\big)\,.\end{aligned}$$

□

All in all, we can conclude that what made the proof of the above admissibility result possible is the convergence of the decrease ratio function

$$\gamma(x,\omega) \coloneqq \frac{F\big(x_+(\omega)\big) - F(x)}{\lambda_{x,\omega}\big(\Delta x(\omega)\big)} = 1 + \frac{\frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 - \rho(x,\omega) - o\big(\big\|\Delta x(\omega)\big\|_X^2\big)}{\big|\lambda_{x,\omega}\big(\Delta x(\omega)\big)\big|} \tag{3.2.39}$$

to something greater equal than one for any $\omega \geq 0$ in the limit of $x \to x_*$. This realization will come to help us out later on when we will discuss the choice of regularization parameters in Section 4.2.

The seemingly paradoxical behavior that full Newton steps yield a better model approximation than Newton steps computed from a modified second order model comes from the fact that f' is not Fréchet differentiable in general. The only prerequisites that we can take advantage of are (3.2.11) and (3.2.24) at fixed x_*.

3.2.6 Alternative Sufficient Decrease Criterion for Numerical Robustness

The result from Proposition 3.2.28 states that our Proximal Newton method as described in Algorithm 9 is well-behaved in the transition from global to local convergence – at least in theory. In addition to the questionable theoretical admissibility of update steps close to optimal solutions, there is yet another peculiarity which might cause trouble in concrete implementations of the method: numerical cancellation. Particularly close to optimal solutions where update steps naturally become very small, the corresponding difference in objective values of subsequent Proximal Newton iterates also deteriorates. As a consequence, computational misbehavior might interfere with the adequate evaluation of algorithmic quantities like the decrease ratio function. Reconsidering the sufficient decrease criterion (3.2.15) given by

$$F\big(x + \Delta x(\omega)\big) - F(x) \leq \gamma \lambda_{x,\omega}\big(\Delta x(\omega)\big)$$

and taking into account that the second order model $\lambda_{x,\omega}$ is naturally less susceptible to numerical cancellation than the direct difference of objective values on the left-hand side, this explains why close to optimal solutions theoretically admissible update steps $\Delta x(\omega)$ are often unjustifiably rejected in practice. As algorithmically intended, this leads to an increase in ω, thus to a shorter update step by Lemma 3.2.26, and thereby again to numerical cancellation. Worst case, this results in the algorithm getting stuck just close before reaching an actual optimal solution of the underlying minimization problem.

Development of Sufficient Decrease Criteria

To evade this inconvenient eventuality, we will consider a different sufficient decrease criterion towards the later stages of the algorithm. With our deliberations made about the original condition from (3.2.15), let us shortly reflect on the role of such criteria and how they can be defined in order to achieve global convergence of the ensuing minimization algorithm:

Generally, sufficient decrease criteria are always intimately related to the globalization mechanism which is used within the respective method: The globalization mechanism has to be able to ensure that the decrease criterion can be fulfilled for sufficiently strong damping/regularization of the step. In the case of line search methods, the step size then has to be sufficiently small while trust region methods restrict the search space to a sufficiently small subset of the original domain. In our scenario, the quadratic norm regularization in (3.2.13) is scaled by a sufficiently large parameter which then ensures satisfiability of the decrease criterion. Having these results in place, the criterion apparently also has to be sufficient for the

proof of global convergence results, mostly the convergence of update step or search direction norms to zero which then again allows to conclude optimality criteria of limit points in some form. Also, for locally accelerated methods, transition results are of importance. In summary, three questions have to be answered for the design of globally convergent algorithms:

(a) Which sufficient decrease criterion is adequate to prove global convergence for the present assumptions on the objective functional and underlying domain space?

(b) Which globalization strategy is able to guarantee satisfiability of this criterion under sufficiently strong damping/regularization?

(c) Does the chosen combination of decrease criterion and globalization strategy allow for unregularized computation of updates close to optimal solutions and thereby unlock local acceleration?

Generally, these questions do not have to be asked anymore for the development of ideas for minimization algorithms since a plethora of intuitive decrease criteria and globalization strategies has already been investigated in the literature. It is, however, reasonable to be aware of the general concepts when developing minimization algorithms. As we have mentioned beforehand, additionally computational robustness and availability of the corresponding algorithmic quantities in concrete implementations play an important role. Let us now define our alternative sufficient decrease criterion and consider to what extent each of the above demands is satisfied by it.

The Alternative Sufficient Decrease Criterion

Computational robustness is ensured by using a first order model in both f and g instead of direct differences within the respective quantities, cf. e.g. [115, Equation (18)] for the case of cubic regularization in a semi-smooth Newton method. For a generalization of this idea, it is intuitive to assess the admissibility of update steps close to stationary points of our problem by an inequality of the form

$$\big[f'\big(x+\Delta x(\omega)\big)+\mu_+\big]\Delta x(\omega) \leq -\frac{1+\gamma}{2}\omega\big\|\Delta x(\omega)\big\|_X^2\,. \tag{3.2.40}$$

Here, $\gamma \in]0,1]$ denotes the sufficient decrease parameter from before and $\mu_+ \in \partial_F g\big(x+\Delta x(\omega)\big)$ is a Fréchet-subderivative of g at the updated iterate. Let us now shortly elaborate on the evaluation of (3.2.40) which also clarifies the particular choice of the latter subderivative. We reconsider optimality conditions of exactly computed update steps in (3.2.13), i.e., the dual space inclusion

$$f'(x)+(H_x+\omega\mathfrak{R})\Delta x(\omega) \in \partial_F g\big(x+\Delta x(\omega)\big)$$

which can be rephrased into the existence of some $\mu_+ \in \partial_F g\big(x+\Delta x(\omega)\big)$ such that

$$\mu_+ = -\big[f'(x)+(H_x+\omega\mathfrak{R})\Delta x(\omega)\big] \tag{3.2.41}$$

holds. This lets us reformulate the alternative sufficient decrease criterion (3.2.40) in a directly computable variant which we use for the corresponding definition of the concept:

Definition 3.2.29: Admissible Regularization Parameters and Update Steps Close to Stationary Points

We say that the regularization parameter $\omega > -\big(\kappa_1(x) + \kappa_2\big)$ and the corresponding update step $\Delta x(\omega)$ computed according to (3.2.13) are **admissible close to stationary points** of (3.2.1) if the **(alternative/numerically robust) sufficient decrease criterion**

$$\big[f'\big(x + \Delta x(\omega)\big) - f'(x) - H_x\big(\Delta x(\omega)\big)\big]\Delta x(\omega) \leq \frac{1-\gamma}{2}\omega\big\|\Delta x(\omega)\big\|_X^2 \tag{3.2.42}$$

holds for the **sufficient decrease parameter** $\gamma \in]0, 1]$.

As we will see over the course of the further investigation of this newly defined sufficient decrease criterion, (3.2.42) is not only superior to (3.2.40) from a computational but also from an analytical standpoint. The formulation via (3.2.40) with the subdifferential element $\mu_+ \in \partial_F g\big(x + \Delta x(\omega)\big)$ is rather of motivational character for using a first order model of F instead of its direct difference on the left-hand side of the sufficient decrease criterion in order to avoid numerical cancellation. For convergence analysis, however, the optimality of our exactly computed update steps $\Delta x(\omega)$ within the subproblem (3.2.13) plays a by far more important role than the subdifferential characterization of μ_+.

From this standpoint, let us now consider satisfiability of (3.2.42) for sufficiently large values of the regularization parameter ω. Again here, as in the proof of Lemma 3.2.12, the Lipschitz-constant L_f of f' plays an important role:

Lemma 3.2.30: Satisfiability of the Alternative Sufficient Decrease Criterion

The alternative sufficient decrease criterion (3.2.40) is satisfied for $\gamma \in]0, 1]$ if the regularization parameter ω suffices the estimate

$$\omega \geq \frac{2(L_f + M)}{1-\gamma}\,.$$

Proof. We consider the equivalent reformulation from (3.2.42) and simply estimate the left-hand side via

$$\big\|f'\big(x + \Delta x(\omega)\big) - f'(x) - H_x\big(\Delta x(\omega)\big)\big\|_{X^*} \leq (L_f + M)\big\|\Delta x(\omega)\big\|_X$$

by the Lipschitz continuity of f' from (A1) and the uniform boundedness of the H_x from (A2). This directly provides us with

$$\big[f'\big(x + \Delta x(\omega)\big) - f'(x) - H_x\big(\Delta x(\omega)\big)\big]\Delta x(\omega) \leq \frac{L_f + M}{\omega}\omega\big\|\Delta x(\omega)\big\|_X^2$$

where we can easily deduce a sufficient lower bound on ω using the prefactor fraction on the right-hand side via

$$\frac{L_f + M}{\omega} \leq \frac{1-\gamma}{2} \quad \Leftrightarrow \quad \omega \geq \frac{2(L_f + M)}{1-\gamma}$$

which concludes the proof of the assertion above. □

With the above guarantee of satisfiability of the alternative sufficient decrease criterion at hand, we can now turn our attention to the demand that global convergence results can be deduced from it in case it is fulfilled. Note here that we employ (3.2.40) only if update steps are already small in some sense, thus we merely need "residual convergence" of the ensuing computation strategy. We will see over the course of the proof what the assumption of update steps being "already small" means in detail.

Proposition 3.2.31: Global Residual Convergence Using the Alternative Decrease Criterion

Let f' be semi-smooth at a stationary point $x_* \in X$ with respect to the mapping $H\colon X \to \mathcal{L}^{(2)}(X,\mathbb{R})$ which is continuous at x_*.

Then, if update steps are already sufficiently small when the alternative sufficient decrease criterion (3.2.40) is used, global convergence from the original algorithm is continued. In particular, we have $\big\|\Delta x_k(\omega)\big\|_X \to 0$ for $k \to \infty$ (in case F is bounded from below) together with all ensuing global convergence results from Section 3.2.3.

Proof. The choice of $\omega > -\big(\kappa_1(x)+\kappa_2\big)$ provides us with (strong) convexity of the regularized second order model $\lambda_{x,\omega}\colon X \to]-\infty,\infty]$ which we use insofar that

$$\lambda_{x,\omega}\big(\Delta x(\omega)\big) \leq -\frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2$$

holds by the optimality of $\Delta x(\omega)$ as first recognized in (3.2.16). In particular, this implies an estimate for the difference in the non-smooth part given by

$$g\big(x+\Delta x(\omega)\big) - g(x) \leq -f'(x)\Delta x(\omega) - \frac{1}{2}\big(H_x+\omega\mathfrak{R}\big)\big(\Delta x(\omega)\big)^2 - \frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2$$

from where we can deduce the following bound for the composite objective functional:

$$\begin{aligned}
&F\big(x+\Delta x(\omega)\big) - F(x)\\
&\quad\leq f\big(x+\Delta x(\omega)\big) - f(x) - f'(x)\Delta x(\omega) - \frac{1}{2}\big(H_x+\omega\mathfrak{R}\big)\big(\Delta x(\omega)\big)^2\\
&\qquad - \frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2\\
&\quad= \big[f\big(x+\Delta x(\omega)\big) - f(x) + \frac{1}{2}H_x\big(\Delta x(\omega)\big)^2\big] - \big[f'(x)+H_x\big(\Delta x(\omega)\big)\big]\Delta x(\omega)\\
&\qquad - \frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 - \frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2\,.
\end{aligned}$$

Let us now turn our attention to the first bracket term. We again use a telescoping strategy similar to the one within the proof of Proposition 3.2.28 and use the corresponding smoothness

assumptions:

$$\begin{aligned}
&f\big(x+\Delta x(\omega)\big) - f(x) + \frac{1}{2}H_x\big(\Delta x(\omega)\big)^2 - f'\big(x+\Delta x(\omega)\big)\Delta x(\omega) \\
&\quad = f\big(x+\Delta x(\omega)\big) - f(x_*) - f'(x_*)\big(x_+(\omega)-x_*\big) - \frac{1}{2}H_{x_+(\omega)}\big(x_+(\omega)-x_*\big)^2 \\
&\qquad - \big[f(x) - f(x_*) - f'(x_*)(x-x_*) - \frac{1}{2}H_x(x-x_*)^2\big] \\
&\qquad + \big[f'(x_*) - f'\big(x+\Delta x(\omega)\big) - H_{x_+}\big(x_*-x_+(\omega)\big)\big]\,\Delta x(\omega) \\
&\qquad + (H_x - H_{x_+(\omega)})\big(x-x_*, \Delta x(\omega)\big) + \frac{1}{2}(H_{x_+(\omega)} - H_x)\big(x_+(\omega)-x_*\big)^2 .
\end{aligned}$$

Firstly, we notice that by Proposition 3.2.22, f is additionally second order semi-smooth at x_* with respect to H. As a consequence, we can estimate the first two terms by the second-order semi-smoothness of f at x_* and the third term with the aid of the semi-smoothness of f' at x_*, both with respect to $H\colon X \to \mathcal{L}^{(2)}(X,\mathbb{R})$. For the last two terms, we take advantage of the continuity assumption on the latter mapping. All in all, the estimates from Corollary 3.2.27 then provide us with

$$f\big(x+\Delta x(\omega)\big) - f(x) + \frac{1}{2}H_x\big(\Delta x(\omega)\big)^2 = f'\big(x+\Delta x(\omega)\big)\Delta x(\omega) + o\big(\big\|\Delta x(\omega)\big\|_X^2\big)$$

in the limit of $x \to x_*$ which we can use in order to find

$$\begin{aligned}
&F\big(x+\Delta x(\omega)\big) - F(x) \\
&\quad \le \big[f'\big(x+\Delta x(\omega)\big) - f'(x) - H_x\big(\Delta x(\omega)\big)\big]\Delta x(\omega) - \frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 \\
&\qquad - \frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2 + o\big(\big\|\Delta x(\omega)\big\|_X^2\big) \\
&\quad \le \frac{1-\gamma}{2}\omega\big\|\Delta x(\omega)\big\|_X^2 - \frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 - \frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2 + o\big(\big\|\Delta x(\omega)\big\|_X^2\big) \\
&\quad = -\frac{\gamma}{2}\omega\big\|\Delta x(\omega)\big\|_X^2 - \frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2 + o\big(\big\|\Delta x(\omega)\big\|_X^2\big) \\
&\quad \le -c\big\|\Delta x(\omega)\big\|_X^2
\end{aligned}$$

again in the limit of $x \to x_*$ by $\Delta x(\omega)$ satisfying the alternative sufficient decrease criterion (3.2.42). The last estimate in the above sequence then stems from the assumption that our update steps $\Delta x(\omega)$ are already sufficiently small close to the stationary point x_* of (3.2.1) where we started to alternatively use (3.2.40) instead of (3.2.15).

The above arguments provide us with a bound of the form (3.2.21) which in particular implies convergence of $\Delta x_k(\omega)$ to zero for $k \to \infty$ in case F is bounded from below just as in Lemma 3.2.13. The ensuing global convergence results can be deduced from first order optimality conditions of (3.2.13) as has been done in Section 3.2.3. □

Remark. *Taking a closer look at the proof from above reveals that we dispense with the continuity assumption on $H\colon X \to \mathcal{L}^{(2)}(X,\mathbb{R})$ in case both the semi-smoothness of f' and second order semi-smoothness of f are demanded at the current iterate $x \in X$ close to x_* instead of the mere semi-smoothness at x_* itself. Under these stronger assumptions, the telescoping argument simplifies and the latter terms involving H evaluated at different points do not appear.*

The above proofs for satisfiability and residual global convergence also reveal the thoughts behind the at first rather irritating choice of the prefactor constant $\frac{1+\gamma}{2} \in]\frac{1}{2}, 1]$ within (3.2.40): For satisfiability, we need the constant to be lower than one while residual global convergence requires it to be any larger than one half. Additionally, tying the criterion to the original one by involving the decrease parameter $\gamma \in]0, 1]$ from before is an intuitive strategy in the choice. Thus, γ also close to optimal controls how restrictive the corresponding sufficient decrease criterion is in rejecting update steps.

Note here that the proof above again has not specifically made use of the subgradient property of μ_+ from (3.2.40) but has rather taken advantage of the optimality of update steps $\Delta x(\omega)$ within their computation subproblem in the form of the norm estimate (3.2.16). This recognition will be crucial when developing a similar strategy in the inexact case later on.

Let us now turn our attention to the investigation to which extent also the alternative sufficient decrease criterion allows for an admissibility result of arbitrarily small regularization parameters like Proposition 3.2.28. At least for any $\omega > 0$ we can achieve a similar result:

Proposition 3.2.32: Admissibility Close to Optimal Solutions Under the Alternative Sufficient Decrease Criterion

Let $f' \colon X \to X^*$ be semi-smooth at an optimal solution $x_* \in X$ of (3.2.1) with respect to the mapping $H \colon X \to \mathcal{L}(X, X^*)$ which is continuous at x_*.

Then, for any $\gamma \in]0, 1]$ and $\omega > 0$ we can find a neighborhood $U_{\gamma,\omega}$ of x_* such that the alternative sufficient decrease criterion (3.2.40) is satisfied by update steps $\Delta x(\omega)$ computed via (3.2.13) at any $x \in U_{\gamma,\omega}$ for that ω.

Proof. Let us consider the left-hand side of the equivalent reformulation (3.2.42) in norm. Similar as in the proof of the original transition result from Proposition 3.2.28, we now telescope the respective operator expression in norm in order to obtain

$$\begin{aligned}
&\big\|f'\big(x + \Delta x(\omega)\big) - f'(x) - H_x\big(\Delta x(\omega)\big)\big\|_{X^*} \\
&\qquad \leq \big\|f'\big(x + \Delta x(\omega)\big) - f'(x_*) - H_{x_*}\big(x + \Delta x(\omega) - x_*\big)\big\|_{X^*} \\
&\qquad\quad + \big\|f'(x_*) - f'(x) + H_{x_*}(x - x_*)\big\|_{X^*} \\
&\qquad\quad + \big\|H_{x_*}\big(x + \Delta x(\omega) - x_*\big) - H_{x_*}(x - x_*) - H_x\big(\Delta x(\omega)\big)\big\|_{X^*} \\
&\qquad = o\big(\big\|\Delta x(\omega)\big\|_X\big) + \big\|(H_{x_*} - H_x)\big(\Delta x(\omega)\big)\big\|_{X^*} = o\big(\big\|\Delta x(\omega)\big\|_X\big)
\end{aligned}$$

in the limit of $x \to x_*$. Here, we used the semi-smoothness of f' with respect to H twice in the second step and for the last identity took advantage of the continuity of the Newton-derivative H. In particular, the equivalences in limit behavior of the occurring quantities for $x \to x_*$ from Corollary 3.2.27 were implicitly important.

This lets us conclude the existence of a modulus of continuity $\psi \colon [0, \infty[\to [0, \infty[$ with $\psi(t) \to 0$ for $t \to 0$ such that we can estimate the left-hand side of (3.2.42) via

$$\big[f'\big(x + \Delta x(\omega)\big) - f'(x) - H_x\big(\Delta x(\omega)\big)\big]\Delta x(\omega) \leq \frac{\psi\big(\big\|\Delta x(\omega)\big\|_X\big)}{\omega}\omega\big\|\Delta x(\omega)\big\|_X^2 .$$

From the global convergence result in this scenario, cf. Proposition 3.2.31, we also here know that update step norms $\big\|\Delta x(\omega)\big\|_X$ tend to zero as we approach the optimal solution of the underlying minimization problem. For any fixed $\omega > 0$, we can thus choose a neighborhood

$U_{\gamma,\omega}$ of x_* such that the modulus of continuity ψ is sufficiently small such that the prefactor fraction also here is smaller than $\frac{1-\gamma}{2} \in [0, \frac{1}{2}[$. This in turn implies admissibility of the corresponding update step. □

Remark. *Also here, we recognize that we can get rid of the continuity assumption on the Newton derivative $H\colon X \to \mathcal{L}(X, X^*)$ if we demand semi-smoothness of f' directly at x instead of only at x_*. This shift of continuity assumptions completely replaces the telescoping argument this time around instead of only simplifying it.*

Remark. *We have to note here that the eventual admissibility of undamped update steps ($\omega = 0$) can not be rigorously verified within the scenario which we have laid out here. Let us shortly discuss a possible modification of* (3.2.40) *in order to guarantee also this desired property of the globalization mechanism: Reconsidering the proofs above, admissibility for $\omega = 0$ would require the choice of some $\gamma_0 > 0$ and the consideration of*

$$\big[f'\big(x + \Delta x(\omega)\big) + \mu_+\big]\Delta x(\omega) \leq -\big(\frac{1+\gamma}{2}\omega - \gamma_0\big)\big\|\Delta x(\omega)\big\|_X^2$$

instead of (3.2.40). *Due to the $o\big(\big\|\Delta x\big\|_X^2\big)$-terms on the right-hand side of*

$$\big[f'(x + \Delta x) + \mu_+\big]\Delta x = \big[f'(x + \Delta x) - f'(x) - H_x(\Delta x)\big]\Delta x = o\big(\big\|\Delta x\big\|_X^2\big)$$

update step norms can then be assumed to be sufficiently small for $x \in X$ close to x_* such that the modified criterion is satisfied also for $\omega = 0$. The above modification by the introduction of γ_0 then on the other hand would hinder the global convergence proof from Proposition 3.2.31 insofar that residual convergence can only be ensured for sufficiently large ω which then again rules out undamped update steps from this perspective. Thus, we decided to stick to our original formulation from (3.2.40).

Before considering the numerically stable reformulation above, we have deduced a similar admissibility result for update steps in Proposition 3.2.28. After its proof, we have concluded that it has been made possible by the convergence of some decrease ratio function $\gamma\colon X \times [0, \omega[\to \mathbb{R}$ to some value greater equal than one in the limit of the current iterate approaching the optimal solution. In a similar fashion, we can also here define a *numerically robust decrease ratio function* via

$$\tilde{\gamma}\colon X \times [0, \omega[\to \mathbb{R}\,, \quad \tilde{\gamma}(x, \omega) \coloneqq \frac{\big[f'\big(x + \Delta x(\omega)\big) - f'(x) - (H_x + \omega\mathfrak{R})\Delta x(\omega)\big]\Delta x(\omega)}{-\omega\big\|\Delta x(\omega)\big\|_X^2} \tag{3.2.43}$$

which has to be larger than the corresponding threshold value $\tilde{\gamma} \coloneqq \frac{1+\gamma}{2} \in]\frac{1}{2}, 1]$ for the update step to be admissible for sufficient decrease according to (3.2.42). From this perspective, our transition result from Proposition 3.2.32 now shows that also this alternative decrease ratio function converges to some value larger equal than one as the sequence of iterates approaches an optimal solution of (3.2.1).

Discussion of Results

Let us shortly address the continuity assumption on the second order Newton derivative $H\colon X \to \mathcal{L}^{(2)}(X,\mathbb{R})$ which we have demanded for the above convergence theory. It removes the concept of second order semi-smoothness from convergence theory in the sense that we can simply deduce the latter from the semi-smoothness of f' together with the continuity of the Newton-derivative H via Proposition 3.2.22.

On that note, even though the results which we have deduced above for the alternative sufficient decrease criterion might at least in parts come over as unsatisfactory in the context, we have to put them in perspective in order to adequately evaluate their quality: Firstly, we have to take into account that we have already formulated a perfectly fine global convergence theory the sufficient decrease criterion of which is tailored to the corresponding globalization method. In theory, this variant also works out arbitrarily close to solutions of the underlying minimization problem but exhibits computational peculiarities in its implementation for concrete examples. From this perspective, it is illustrative to think of algorithmic functionality as an interplay of rigorous theory which is formulated in sufficient generality and its implementation which is different from case to case and might be affected by computational limitations. While it is always better to stay close to assumptions and ensuing guaranteed results from theory, sometimes one has to consider the trade-off in favor of rather heuristic and more robust approaches.

Secondly, the generality in which our theoretical framework has been formulated also affects the quantifiability of errors and limit expressions: Replacing semi-smoothness of f' and second order semi-smoothness of f only at stationary points with more generous differentiability properties of at least the smooth part f everywhere in X would provide us with an easier way of evaluating such quantities. For example, assuming second order differentiability with a Lipschitz hessian gives rise to direct estimates using cubic norm terms with the corresponding Lipschitz constant as a prefactor. This then again allows for a direct treatment of the respective remainder terms. While the ensuing transparency in formulation seems desirable, pursuing generality with respect to differentiability assumptions is superior both from an analytical and an application-focused standpoint.

Algorithmic Strategy

Let us now use the convergence theory developed above in order to create a modification of the scheme from Algorithm 9 which is able to adequately cope with numerical instability of computation close to optimal solutions of (3.2.1). Due to the drawbacks of the alternative sufficient decrease criterion in view of convergence theory which we have discussed above, we will try to avoid the use of (3.2.42) as long as we can in favor of the original formulation (3.2.15). Once we are close to solutions of (3.2.1), however, we will use (3.2.42) in order to avoid numerical cancellation as follows:

We characterize "being close to a solution" by an additional *proximity criterion* in the form of a norm bound

$$(1+\omega)\big\|\Delta x(\omega)\big\|_X \leq \tilde{\varepsilon} \tag{3.2.44}$$

which mimics the stopping criterion from Algorithm 9 but uses a larger threshold value $\tilde{\varepsilon} > \varepsilon > 0$. Once (3.2.44) holds, we first test the conventional form (3.2.15). In case it fails, we will then make sure that this did not happen due to numerical instability. For this reason, we will

not as before increase the regularization parameter and recompute the update step right away but first additionally test the alternative criterion (3.2.42). If the considered update suffices the latter, we can take advantage of the global convergence theory deduced above and use the step. Otherwise, we increase ω and recompute $\Delta x(\omega)$ as usual.

Furthermore, due to numerical instability, we also let go of the bound (3.2.21) once the proximity criterion (3.2.44) holds. Global convergence is from there on taken care of adequately by the alternative sufficient decrease criterion.

We summarize the algorithmic strategy of this modified form of Algorithm 9 in the scheme of Algorithm 10. For a more illustrative overview, we refer to the algorithmic conclusion in Section 4.3 where the final form of the algorithm is presented in Figure 4.11.

Algorithm 10: Second Order Semi-smooth Proximal Newton Method Modified for Numerical Stability

Data: Starting point $x_0 \in X$, sufficient decrease parameter $\gamma \in]0, 1]$, initial value $\omega_0 \geq 0$, thresholds $0 < \tilde{\varepsilon} < \varepsilon$ for the stopping and proximity criterion
Initialization: $k = 0$;
while $(1 + \omega_k)\left\|\Delta x_k(\omega_k)\right\|_X \geq \varepsilon$ **do**
  Compute a trial step $\Delta x_k(\omega_k)$ according to (3.2.13);
  if $(1 + \omega_k)\left\|\Delta x_k(\omega_k)\right\|_X \geq \tilde{\varepsilon}$ **then**
    while *bound* (3.2.21) *or sufficient descent criterion* (3.2.15) *is not satisfied* **do**
      Increase regularization parameter ω_k adequately;
      Recompute trial $\Delta x_k(\omega_k)$ step as above;
    end
  else
    while *sufficient descent criterion* (3.2.15) *and alternative version* (3.2.42) *are not satisfied* **do**
      Increase regularization parameter ω_k adequately;
      Recompute trial $\Delta x_k(\omega_k)$ step as above;
    end
  end
  Update the current iterate to $x_{k+1} \leftarrow x_k + \Delta x_k(\omega_k)$;
  Decrease ω_k appropriately to some $\omega_{k+1} < \omega_k$ for next iteration;
  Update the sequence index $k \leftarrow k + 1$;
end

3.2.7 Numerical Results

The Proximal Newton method which we have developed above constitutes a working minimization procedure which is applicable to function space problems. Even though there are still modifications to be made with regard to algorithmic efficiency, it is illustrative to showcase the functionality of the method at this early stage and compare it to existing strategies.

The Objective Functional

To this end, we consider the following problem on $\Omega = [0,1]^2 \subset \mathbb{R}^2$: Find $u \in H_0^1(\Omega)$ that minimizes the composite objective functional $F\colon H_0^1(\Omega) \to \mathbb{R}$ defined via

$$F(u) \coloneqq \int_\Omega \frac{1}{2}\big\|\nabla u\big\|_{\mathbb{R}^2}^2 + \alpha \max\{\big\|\nabla u\big\|_{\mathbb{R}^2} - 1, 0\}^2 + \beta u^3 + c\,|u| + \rho u \,\mathrm{d}x\,. \tag{3.2.45}$$

with parameters $c > 0$ and $\alpha, \beta \in \mathbb{R}$ as well as a force field $\rho : \Omega \to \mathbb{R}$. The norm $\|\cdot\|_{\mathbb{R}^2}$ denotes the Euclidean 2-norm on $\mathbb{R}^2$. In the sense of the theory of the preceding sections we can identify the smooth part of F as $f : H_0^1(\Omega) \to \mathbb{R}$ given by

$$f(u) \coloneqq \int_\Omega \frac{1}{2}\big\|\nabla u\big\|_{\mathbb{R}^2}^2 + \alpha \max\{\big\|\nabla u\big\|_{\mathbb{R}^2} - 1, 0\}^2 + \beta u^3 + \rho u \,\mathrm{d}x\,.$$

We have to note here that f technically does not satisfy the assumptions made on the smooth part of the composite objective functional specified above in the case $\alpha \neq 0$ due to the lack of semi-smoothness of the corresponding squared max-term. The use of the derivative ∇u instead of function values u creates a norm-gap which can not be, as usual, compensated by Sobolev-embeddings and hinders the proof of semi-smoothness of the respective superposition operator. However, we think that slightly going beyond the framework of theoretical results for numerical investigations can be instructive.

For our implementation of the solution algorithm we chose the force field ρ to be constant on its domain and equal to some so called load-factor $\tilde{\rho} > 0$ which we will from now on refer to as simply ρ. Consequently, the non-smooth part of the objective functional g only consists of the scaled integral over the absolute value term which apparently also satisfies the specifications made on g before. Note that the underlying Hilbert space is given by $X = H_0^1(\Omega, \mathbb{R})$ which also determines the norm choice for regularization of the subproblem.

Specifics of the Implementation

In the following, we will dive deeper into the specifics of our implementation of the algorithm: In order to differentiate the smooth part of the composite objective functional and create a second order model of it around some current iterate, we take advantage of the automatic differentiation software package adol-C, cf. [114]. With the second order model at hand, we can then consider subproblem (3.2.13) which has to be solved in order to obtain a candidate for the update of the current iterate. As mentioned beforehand, for the latter endeavor we employ the Truncated Non-smooth Newton MultiGrid (TNNMG) method, cf. Appendix Section A.1. Roughly speaking, we can summarize this method as a mixture of exact, non-smooth Gauß-Seidel steps for each component and global, truncated Newton steps enhanced with a line-search procedure. The stopping criterion for the minimization process within TNNMG is given by a relative norm threshold for increments and the scheme is analytically proven to converge for convex and coercitive problems, cf. [33]. For detailed information on the test machine which we have used in order to conduct our tests, we refer to Appendix Section A.3.

However, the most delicate issue concerning the implementation of our algorithm and its application to the problem described above is the choice of the regularization parameter $\omega \geq 0$ along the sequence of iterates $(x_k) \subset X$. For now, we will confine ourselves to displaying the convergence properties of the class of Proximal Newton methods in the scenario presented above and not attach too much value to algorithmic technicalities. After setting it to the initial value $\omega_0 = 64$, we take the rather heuristic approach of simply doubling ω in case the

respective sufficient decrease criterion (3.2.15) or (3.2.40) (for $\gamma = \frac{1}{2}$) is not satisfied by the current update step candidate. If the update is accepted, we multiply ω by $\frac{1}{2^n}$ where $n \in \mathbb{N}$ denotes the number of consecutive accepted update steps. The latter feature ensures that the region of local acceleration is recognized by the algorithm and the regularization parameter then quickly decreases once the iterates come close to the minimizer. For the superlinear convergence from Theorem 3.2.8 to arise, undamped update steps have to be conducted, i.e., the regularization parameter has to be zero and not merely sufficiently small. We will see later that ω converging to zero actually suffices in that regard which is why we will stick to the aforementioned approach for now.

Even though the choice of ω considered here is rather heuristic and not problem-specific at all, it stands in conformity with the theory established over the course of the previous sections. In addition, it successfully displays the global convergence and local acceleration of our Proximal Newton method for the model problem of minimizing (3.2.45) over $H_0^1(\Omega)$.

As far as stopping criteria of the Proximal Newton method are concerned, we chose $\varepsilon = 10^{-10}$ in Algorithm 9 which yields

$$(1+\omega)\big\|\Delta x(\omega)\big\|_X \leq 10^{-10} \tag{3.2.46}$$

as the crucial requirement of ending the computation of update steps. The proximity criterion (3.2.44), which determines the additional use of the alternative sufficient decrease criterion from Section 3.2.6, is closely related to this stopping criterion. For our numerical test here, we choose the respective threshold value as $\tilde{\varepsilon} = 10^{-4}$. Furthermore, the constant determining the bound (3.2.21) is set to $\overline{M} = 10^{10}$.

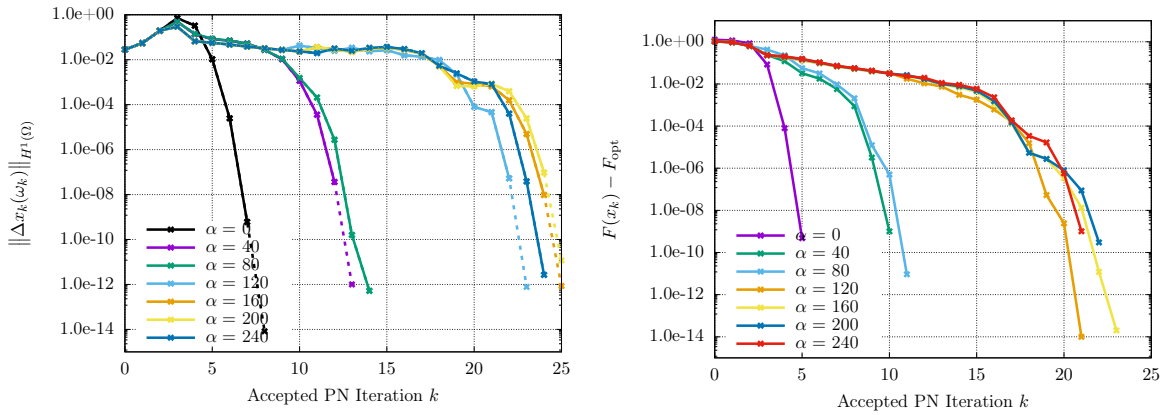
Test Scenarios and Results

For our numerical investigations, we fix the parameters $\beta = c = 10$ and $\rho = -20$ in the same order of magnitude and vary the influence of the squared max norm term by increasing the corresponding prefactor $\alpha \in \{0, 40, ..., 240\}$. We conduct five uniform grid refinements of the domain Ω which results in $4^6 = 4096$ grid elements.

The norm of update steps $\Delta x(\omega)$ is depicted in Figure 3.1a. This illustration not only highlights the local superlinear convergence of our method but also shows that, as α increases, the corresponding minimization problem becomes significantly harder to solve. In particular, the globalization phase and thereby number of Proximal Newton iterations required for finding the solution grows significantly across this series of tests.

Note that for some computational scenarios the graph of update step norms does not actually reach the threshold value $\varepsilon = 10^{-10}$. This fact can be attributed to the last step not satisfying the respective sufficient decrease criterion. Our measure for optimality of the current iterate, however, does not depend on the update step being accepted by the globalization procedure but only assesses the norm of an update step as an estimate for the distance of the Fréchet subdifferential $\partial_F F$ to zero in X^*, cf. (3.2.22). For this reason, we stop update step computation also in this case. For the convenience of the reader, we also added these last declined update steps as a dashed extension of the respective plot of correction norms.

Another meaningful quantity the behavior of which signifies the local superlinear convergence of our method is the objective function value. From a physical point of view, the objective value is sometimes also referred to as the energy which has to be minimized for the solution of the respective application problem. A plot for the energy difference to the respective optimal value F_{opt} across the minimization process of our test scenarios is depicted



(a) Correction Norms $\|\Delta x_k(\omega_k)\|_{H^1(\Omega)}$.

(b) Energy Differences $F(x_k) - F_{\text{opt}}$.

Figure 3.1: Graphs of correction norms and energy differences to the optimal value for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$ for the Proximal Newton method. Dashed extensions show the last declined update.

in Figure 3.1b. Note here that the plot only incorporates iteration indices up to the second to last ones since we assume the last objective value to be optimal. The difference to this value would hence be equal to zero which can not be depicted in a logarithmic fashion.

Furthermore, also the transition result from Section 3.2.5 can be visualized within the numerical investigation of our exact Proximal Newton method. In that regard, consider Figure 3.2 the plot of regularization parameters employed within the second order model $\lambda_{x,\omega}$ from (3.2.12) in order to compute damped update steps according to (3.2.13). Note here that the plot not only shows the successful Proximal Newton iterations but also incorporates declined update steps which gives a better insight into the globalization phase of our algorithm. Additionally, it is apparent that – close to the solution of our problem – arbitrarily small regularization parameters lead to admissible update steps for our sufficient decrease criteria. This enables the local accelerated convergence which we have recognized across the previously described figures.

Lastly, we investigate the mesh-independence of our Proximal Newton method the convergence theory for which has been formulated in general function space. For that reason, the convergence behavior of our method should not depend on the number of grid refinements which we conduct prior to solving the discretized problem. The number of Proximal Newton steps required for finding a solution across all of our testing scenarios for different mesh sizes h can be retraced in Table 3.1.

Taking a closer look reveals that the number of iterations stays constant in the case of $\alpha = 0$, i.e., we have the desired mesh independence result in the scenario which actually fits into our theoretical framework from above. As α grows across the rest of the test series, we recognize a slight increase in the number of required update steps in order to find the solution. This fact can be ascribed to the lacking semi-smoothness of the squared max norm term with gradient norms the influence of which becomes larger as α is increased.

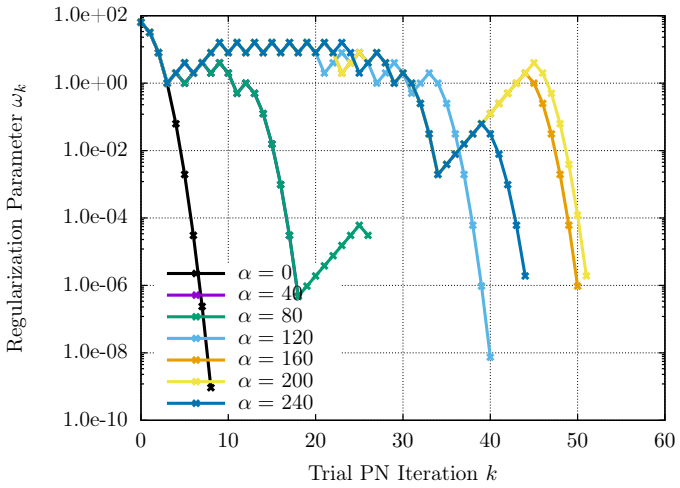


Figure 3.2: Regularization Parameters ω_k within update step computation for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$.

Comparison to Proximal Gradient and FISTA

Let us now consider other standard methods which can be used in order to solve composite minimization problems and compare their performance in minimizing the objective functional from above to our findings for the Proximal Newton method. To this end, we consider both the standard Proximal Gradient method and its accelerated variant FISTA as we have presented them over the course of our introductory elaborations in Section 3.1.3.

The test with specifications as made above for the Proximal Newton test series immediately exposed a general weakness of both the Proximal Gradient and FISTA method as presented here: Even for the rather simple scenarios of $\alpha \in \{0, 40\}$, both methods could not reach the prescribed accuracy in the form of the update step norm stopping criterion from (3.2.46) but always got stuck close to that mark due to numerical instability of the respective sufficient decrease criterion. As a consequence both methods tried to increase the regularization parameter further and further until a prescribed emergency stop threshold value has been reached. In parts, this has also been the case for the Proximal Newton method prior to the introduction of our numerically robust alternative sufficient decrease criterion. This shows that our adaptations for numerical robustness from Section 3.2.6 pay off and we do not have to bother with such peculiarities thanks to our algorithmic deliberations.

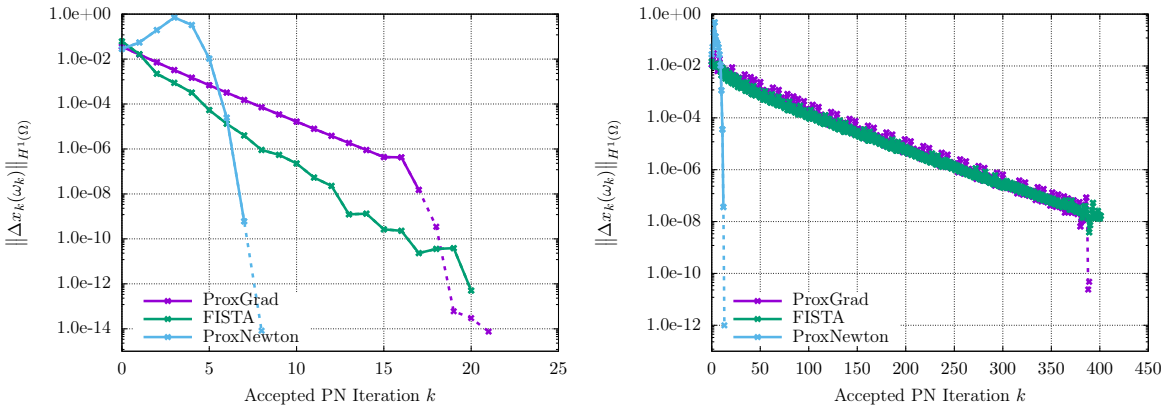
However, our Proximal Newton method is superior to the corresponding first order alternatives not only regarding general convergence behavior. Also the speed of achieving a sufficiently accurate solution speaks for itself. Figure 3.3 shows the correction norms of Proximal Gradient ("ProxGrad"), FISTA, and Proximal Newton ("ProxNewton") for the rather simple test scenarios of $\alpha \in \{0, 40\}$. For better comparability, we have measured the average computation wall-time for one accepted iteration of each method. As a consequence, we count multiple first order iterations as one meta-iteration of Proximal Newton and display the corresponding ratio of points in plots accordingly.

In the test with $\alpha = 0$, this average computational time was $\bar{t} \approx 0.106s$ for FISTA and Proximal Gradient, and $\bar{t} \approx 0.669s$ for Proximal Newton which – rounding up – yields a number of seven first order iterations as one meta-iteration for our second order method. For $\alpha = 40$, we have $\bar{t} \approx 0.107$ for the first order variants and $\bar{t} \approx 0.917s$ for Proximal Newton. Thus, we

h \ α	0	40	80	120	160	200	240
2^{-4}	9	12	12	17	18	18	20
2^{-5}	9	14	15	19	20	18	23
2^{-6}	8	13	15	23	25	25	25
2^{-7}	9	14	19	25	28	29	29
2^{-8}	9	16	21	30	31	33	32

Table 3.1: Number of accepted iterations N for different grid sizes h and prefactor values α for fixed parameters $\beta = 10$, $c = 10$ and $\rho = -20$.

count nine first order iterations as one meta-iteration for Proximal Newton. The dashed lines within the plots for Proximal Gradient and FISTA mark the update steps resulting from strong increases of the regularization parameter towards the end of the algorithm shortly before it has been stopped due to the emergency stopping criterion for too strong regularization. For the Proximal Newton graph, the dashed line as before represents the last update step, which has not been accepted, but is sufficiently small in norm for the computation to be stopped.



(a) Correction Norms for $\alpha = 0$.

(b) Correction Norms for $\alpha = 40$.

Figure 3.3: Comparison of Proximal Newton to its first order alternatives. For Proximal Newton, dashed lines again represent the last declined update leading to stopping the algorithm. For Proximal Gradient and FISTA, they show updates which result from ramping regularization due to numerical instability.

Even under these circumstances, it is apparent that our Proximal Newton method outperforms the first order alternatives also in view of convergence speed. Adding more non-linearity by increasing the α–prefactor in (3.2.45) makes these differences in efficiency even more significant and aggravates the solvability of the presented minimization problem for first order

approaches even more. Furthermore, we recognize that while the accelerated convergence of FISTA in comparison with Proximal Gradient is visible for the rather smooth scenario of $\alpha = 0$ in Figure 3.3a, this effect deteriorates in the more involved case depicted in Figure 3.3b.

As a consequence of the poor performance of first order alternatives even for the simple function space problem which we have investigated here, we decided to avoid comparisons of our Proximal Newton method with both Proximal Gradient and FISTA for the significantly more demanding problems which are yet to come in the remainder of the present treatise.

All in all, we can see that the "raw" Proximal Newton method as we have presented it over the course of the current chapter constitutes a working and efficient solution method for composite minimization problems of the form (3.2.1). In what follows, we will turn our attention to algorithmic modifications which enhance our minimization strategy with even more computational efficiency on top of the functionality ensured up until now.

Chapter 4

Modifications for Algorithmic Efficiency

As we have explained in the introductory part of the previous chapter, the emphasis there was put on algorithmic functionality, i.e., on the introduction of a reliable minimization algorithm for function space problems of the form (3.2.1) which can handle weakened convexity and differentiability assumptions also in a function space setting and exhibits satisfying convergence properties. With such a method at hand, we can now shift this focus towards algorithmic efficiency of our computational procedure. One intuitive idea which appears promising in that regard has already been addressed when considering Newton methods towards the end of Section 3.1.1: inexact solution of update step subproblems. It is known that there – with the aid of adequately defined *inexactness criteria* – a considerable amount of computational effort can be saved while still preserving the advantageous convergence properties of the second order method both globally and locally.

On a different note, another possibility to gain algorithmic efficiency has also already struck our eye when we described the concrete implementation of our method for the numerical investigations of Section 3.2.7: the choice of the regularization parameter ω in (3.2.13). In that regard, it is particularly interesting to study adaptive approaches to the choice of algorithmic parameters. This field of research has been thoroughly considered in the literature which puts us in the favorable position of being able to get inspiration from previous works and try to generalize some existing ideas to our Proximal Newton setting.

Chapter Outline

In order to address both of the topics mentioned above, we pursue the following structure within the current chapter: We start with Section 4.1 where we introduce the notion of inexactness to the determination of update steps by providing inexactness criteria which are tailored to the function space scenario we find ourselves in, reduce computational effort significantly and still preserve the convergence properties of the exact method. In Section 4.2, we then introduce adaptive strategies for the choice of algorithmic parameters, yet again enhancing both the convergence behavior and the robustness of our algorithm with respect to the application at hand. At last, we present an algorithmic conclusion of the final form of our then inexact Proximal Newton method in Section 4.3 where the main features of our algorithm can be retraced without having to dig through pages of convergence analysis.

4.1 Inexact Computation of Update Steps

As pointed out beforehand, the goal is to carry over the results considering inexact computation of update steps, which we have gathered for both smooth and Proximal Newton Methods in Euclidean space in Section 3.1, to our generalized variant of Proximal Newton methods here. Before departing on this endeavor, let us give a short overview of proceedings in this direction in the (recent) literature.

Proceedings in Recent Literature

The use of gradient-like inexactness criteria, which can be seen as the direct generalization of the one for classical smooth Newton methods in [19] (or (3.1.10) here), is quite common, cf. [11, 44, 55]. A possible form of this generalization – similar to the one considered in [55] – has been introduced in Section 3.1.3 via (3.1.22). There, we have mentioned that the additional knowledge of bounds on the second-order bilinear forms is necessary and that the infinite dimensionality of our minimization framework hinders the evaluation of such criteria. We will go into detail on this peculiarity later. Additionally, the Lipschitz constant of f' must be accessible for the corresponding choice of forcing terms and only local convergence has been investigated in the inexact case.

Globalization of the ensuing method has been achieved in [44] by using a Proximal Gradient substitute step in case the inexactly computed second order step does not suffice a sufficient decrease criterion or the step computation subproblem is ill-formed due to non-convexity which thus can be overcome as well. As we will explain more elaborately later on, this strategy of resorting to first order updates as well is not an as viable option in a function space setting as it is in the Euclidean $\mathbb{R}^n$-scenario. Our solution to the need of globalization via inexact update steps will be the introduction of a second criterion which these increments will have to satisfy.

In [11], the particular case of L_1-regularization for machine learning applications is considered. For this reason, the inexactness criterion from (3.1.22) is further specified and enhanced with a decrease criterion in the quadratic approximation of the composite objective function. The latter is then tightened over the course of the algorithm in order to achieve local acceleration.

Another approach to inexactness criteria is measuring the residual within the step computation subproblem. In [56], where objective functions consisting of the sum of a thrice continuously differentiable and self-concordant smooth part and a convex non-smooth part are considered, the residual vector within optimality conditions for update computation is supposed to be bounded in norm with respect to the already computed inexact step. However, the residual can also be measured via functional descent in the quadratic approximation of the composite objective F, cf. [53, 96]. While in [53] the second order model decrease bound against its optimal value is not directly tested but simply assumed to hold after a finite (and fixed) number of subproblem solver iterations, the authors in [96] take the structure of their randomized coordinate descent subproblem solver into account and also give quadratic bounds for the prefactor constant within their model descent estimate in order to obtain sufficient convergence results.

As has already been the case in the development of general Proximal Newton methods in Section 3.1.3, all of the above works have in common that they depend on the finite dimensional structure of the underlying Euclidean space. In particular, the efficient computation of proximal gradients, required for the evaluation of most inexactness criteria, relies on the

diagonal structure of the underlying scalar product $\langle \cdot, \cdot \rangle_X$, which is usually not present in (discretized) function spaces. Moreover, all current approaches consider fixed search directions which are then scaled by some step length parameter.

Contributions and Assumptions

Our contributions described within this section beyond their work can be summarized as follows: Most importantly, we again replace the Euclidean space setting with a Hilbert space one in order to rigorously allow for function space applications of our method. In particular, we are interested in the important case where X is a Sobolev space. Then, as previously mentioned a diagonal approximation of $\langle \cdot, \cdot \rangle_X$ after discretization would lead to proximal operators that suffer from mesh-dependent condition numbers. Consequently, our inexactness criteria need to be constructed in such a way that their evaluation is efficient in this context. Existing criteria can only be employed efficiently, if $\langle \cdot, \cdot \rangle_X$ enjoys a diagonal structure.

In order to develop a mere augmentation of our already developed Proximal Newton method, the assumptions which we make on the underlying domain space and the respective parts of the composite objective functional from (3.2.1) remain unchanged from the ones described in the introductory part of Section 3.2. In particular, the standing assumptions (A1)-(A4) pertain to hold across the whole of the sections. Our elaborations here are closely related to the ones of the submitted preprint [83].

Section Outline

Let us now briefly outline the structure of the current section: At first, we will introduce the notion of composite gradient mappings and consider some of their basic properties in Section 4.1.1. Afterwards, in Section 4.1.2, we take advantage of the acquired knowledge and introduce the first inexactness criterion in order to investigate local convergence of our method. In addition, we will be able to quantify the influence of both damping and inexactness on the local convergence rate. Section 4.1.3 then considers the globalization phase of our inexact Proximal Newton method and for this reason introduces a second inexactness criterion which compares the functional decrease of inexact updates with steps originating from a simpler subproblem. Thus, we also here achieve sufficient global convergence results. In order to then benefit from local acceleration, we once again investigate the transition to local convergence in Section 4.1.4 in the inexact scenario. As before, to this end we need to ensure that close to optimal solutions also arbitrarily weakly damped update steps yield sufficient decrease. The criterion for the latter is then again slightly modified in favor of numerical robustness in Section 4.1.5. Lastly, we put our method to the test in Section 4.1.6 and display the increased computational efficiency with unchanged algorithmic functionality considering a more demanding model problem in function space.

4.1.1 Composite Gradient Mappings and Their Properties

The main goal to keep in mind is not only to introduce the concept of inexactness to the computation of update steps for the Proximal Newton method from Algorithm 9 but also to quantify the influence of damping update steps to the local convergence rate of said algorithm.

Definition and Representation via Proximal Mappings

For this cause, we take advantage of the notion of so-called *regularized composite gradient mappings* which have already played a minor role within our introductory section 3.1.3 for the generalization of classical inexactness criteria to the composite framework of Proximal Newton methods. For us here, however, they are of crucial importance not for the definition of inexactness criteria or quantifiable algorithmic quantities but only for convergence analysis "behind the scenes". Obviously, we also generalize their definition from (3.1.18) to our Hilbert space setting opposing the finite dimensional one from before.

Adopting a similar notational structure, we introduce $G_\tau^\Phi : X \to X$ for some composite functional $\Phi : X \to]-\infty, \infty]$ of the form $\Phi(x) \coloneqq \phi(x) + \psi(x)$ with smooth part $\phi : X \to \mathbb{R}$ and non-smooth part $\psi : X \to]-\infty, \infty]$. The aforementioned mapping is defined via

$$G_\tau^\Phi(y) \coloneqq -\tau \left[\operatorname*{arg\,min}_{\delta y \in X} \phi'(y)\delta y + \frac{\tau}{2}\big\|\delta y\big\|_X^2 + \psi(y+\delta y) - \psi(y) \right] \tag{4.1.1}$$

for $y \in X$ and some regularization parameter $\tau > 0$ the assumptions on which we will specify over the course of the current section. For the derivation of useful estimates for composite gradient mappings, we remember the definition of scaled dual proximal mappings $\mathcal{P}_\psi^H : X^* \to X$ from (3.2.5), given by

$$\mathcal{P}_\psi^H(\varphi) \coloneqq \operatorname*{arg\,min}_{z \in X} \psi(z) + \frac{1}{2}H(z)^2 - \varphi(z)$$

for arbitrary $\varphi \in X^*$ and some symmetric bilinear form H sufficing (A3) as well as some real valued function ψ satisfying (A4) for constants $\kappa_1, \kappa_2 \in \mathbb{R}$ with $\kappa_1 + \kappa_2 > 0$.

With the aid of scaled proximal mappings, we can express the composite gradient mapping from (4.1.1) as

$$G_\tau^\Phi(y) = \tau\big[y - \mathcal{P}_\psi^{\tau\mathfrak{R}}\big(\tau\mathfrak{R}y - \phi'(y)\big)\big]\,. \tag{4.1.2}$$

This representation enables us to take advantage of our results concerning scaled proximal mappings also when investigating gradient mappings, namely the general estimate from Proposition 3.2.3 and the Lipschitz continuity result from Corollary 3.2.4.

Let us now justify the designation of G_τ^Φ as a regularized composite gradient mapping. If we consider the smooth case of $\psi = 0$ in (4.1.1), the proximal mapping takes the form $\mathcal{P}_\psi^H(\varphi) = H^{-1}\varphi$. This fact carries over to the definition of the gradient mapping via

$$G_\tau^\phi(y) = \tau\big[y - (\tau\mathfrak{R})^{-1}\big(\tau\mathfrak{R}y - \phi'(y)\big)\big] = \mathfrak{R}^{-1}\phi'(y)$$

which resembles the infinite dimensional counterpart of the gradient $\nabla\phi$ in Euclidean space. Note that this consistency result holds for all $\tau > 0$.

Another consideration which expresses the consistency between G_τ^F and some actual smooth gradient of $F = f + g$ with respect to our minimization problem (3.2.1) is the following: Let then $G_\tau^F(x_*) = 0$ hold for some $x_* \in X$ and $\tau \geq -\kappa_2$. This is equivalent to the fixed point equation $x_* = \mathcal{P}_g^{\tau\mathfrak{R}}\big(\tau\mathfrak{R}x_* - f'(x_*)\big)$ which can then again be transformed to $-f'(x_*) \in \partial_F g(x_*)$ in X^*. Consequently, we recognize that the composite gradient mapping is zero if and only if we evaluate it at stationary points of the underlying minimization problem (3.2.1). For

this reason, the norm of the gradient mapping at some iterate is often taken as a measure for optimality and its convergence to zero suffices to prove global convergence results of the ensuing method by first order conditions of the corresponding subproblem. However, our global convergence proof here will take another route.

Key Properties and Auxiliary Estimates

For now, let us derive some key properties of composite gradient mappings which will come in handy as we quantify the influence of both inexactness and damping to local convergence rates of our algorithm.

Before departing on this endeavor we introduce the modified quadratic model $\hat{F}_{x,\omega} : X \to]-\infty, \infty]$ of the composite objective functional F around $x \in X$ with regularization parameter $\omega \geq 0$ via

$$\begin{aligned} \hat{F}_{x,\omega}(y) &\coloneqq F(x) + \lambda_{x,\omega}(y-x) \\ &= f(x) + f'(x)(y-x) + \frac{1}{2}H_x(y-x)^2 + g(y) + \frac{\omega}{2}\|y-x\|_X^2 . \end{aligned} \tag{4.1.3}$$

The corresponding composite gradient mapping $G_\tau^{\hat{F}_{x,\omega}}$ will play an important role. In that regard, we note that in the framework of the definition of the gradient mapping in (4.1.1) we thus have $\hat{F}_{x,\omega} = \Phi = \phi + \psi$ with

$$\phi(y) = f(x) + f'(x)(y-x) + \frac{1}{2}\big(H_x + \omega\mathfrak{R}\big)(y-x)^2 \quad \text{and} \quad \psi(y) = g(y) \tag{4.1.4}$$

and thereby $\phi'(y) = f'(x) + \big(H_x + \omega\mathfrak{R}\big)(y-x)$ for any $y \in X$. The following lemma provides us with helpful estimates for the norm difference of the corresponding composite gradient mapping images both from above and below:

Lemma 4.1.1: Regularity of the Composite Gradient Mapping

For every $x, y, z \in X$ and the choice $\tau \coloneqq \omega + \frac{1}{2}\big(\|H_x\|_{\mathcal{L}(X,X^*)} + \kappa_1(x)\big)$, the regularized composite gradient mapping suffices the estimates

$$\tau\big(1-\mathcal{H}\big)\|y-z\|_X \leq \|G_\tau^{\hat{F}_{x,\omega}}(y) - G_\tau^{\hat{F}_{x,\omega}}(z)\|_X \leq \tau\big(1+\mathcal{H}\big)\|y-z\|_X \tag{4.1.5}$$

where we have abbreviated the constant expression $\mathcal{H} \coloneqq \frac{\|H_x\|_{\mathcal{L}(X,X^*)} - \kappa_1(x)}{2(\tau+\kappa_2)}$.

Proof. As we insert the characterizations of the respective regularized composite gradient mappings as in (4.1.2), we perceive that we can represent their norm difference via

$$\|G_\tau^{\hat{F}_{x,\omega}}(y) - G_\tau^{\hat{F}_{x,\omega}}(z)\|_X = \tau\|(y-z) - \big(\mathcal{P}_y - \mathcal{P}_z\big)\|_X$$

with abbreviations $\mathcal{P}_\xi \coloneqq \mathcal{P}_g^{\tau\mathfrak{R}}\big(\tau\mathfrak{R}\xi - \big[f'(x) + \big(H_x + \omega\mathfrak{R}\big)(\xi - x)\big]\big)$ for $\xi \in \{y, z\}$. Apparently, this provides us with the bounds

$$\tau\big(\|y-z\|_X - \|\mathcal{P}_y - \mathcal{P}_z\|_X\big) \leq \|G_\tau^{\hat{F}_{x,\omega}}(y) - G_\tau^{\hat{F}_{x,\omega}}(z)\|_X \leq \tau\big(\|y-z\|_X + \|\mathcal{P}_y - \mathcal{P}_z\|_X\big)$$

from above and below for the norm difference of gradient mappings. This shows that for the proof of (4.1.5) it suffices to verify

$$\big\|\mathcal{P}_y - \mathcal{P}_z\big\|_X \leq \mathcal{H}\big\|y-z\big\|_X = \frac{\big\|H_x\big\|_{\mathcal{L}(X,X^*)} - \kappa_1(x)}{2(\tau+\kappa_2)}\big\|y-z\big\|_X\,. \tag{4.1.6}$$

The Lipschitz result from Proposition 3.2.3 allows us to establish the following estimate for the norm difference of proximal mapping images in relation to their arguments:

$$\begin{aligned}\big\|\mathcal{P}_y - \mathcal{P}_z\big\|_X &\leq \frac{1}{\tau+\kappa_2}\big\|\tau\mathfrak{R}y - \big(H_x + \omega\mathfrak{R}\big)(y-x) - \big(\tau\mathfrak{R}z - \big(H_x+\omega\mathfrak{R}\big)(z-x)\big)\big\|_{X^*}\\ &= \frac{1}{\tau+\kappa_2}\big\|\big((\tau-\omega)\mathfrak{R} - H_x\big)(y-z)\big\|_{X^*}\\ &\leq \frac{\big\|(\tau-\omega)\mathfrak{R} - H_x\big\|_{\mathcal{L}(X,X^*)}}{\tau+\kappa_2}\big\|y-z\big\|_X\,.\end{aligned} \tag{4.1.7}$$

Let us now pay particular attention to the $\mathcal{L}(X,X^*)$-norm difference in the prefactor above. On the one hand, for any $\tau > -\kappa_2$ ensuring well-definedness of the gradient mapping, we can estimate it by

$$\big\|(\tau-\omega)\mathfrak{R} - H_x\big\|_{\mathcal{L}(X,X^*)} \leq |\tau-\omega| + \big\|H_x\big\|_{\mathcal{L}(X,X^*)}\,.$$

Nevertheless, with further assumptions on the gradient mapping regularization parameter τ we can deduce a better bound. To this end, we define $\zeta \coloneqq \tau - \omega$ and choose ζ_{opt} such that $\big\|\zeta_{\mathrm{opt}}\mathfrak{R} - H_x\big\|_{\mathcal{L}(X,X^*)}$ is minimal. It is easy to see that the eigenvalues of the self-adjoint operator $H_x^\tau \coloneqq \mathfrak{R}^{-1}(\zeta\mathfrak{R} - H_x)$ lie in the interval $\big[\zeta - \big\|H_x\big\|_{\mathcal{L}(X,X^*)}, \zeta - \kappa_1(x)\big]$.

In order to now minimize the norm of H_x^τ, we recognize that it equals the spectral radius of H_x^τ and thus want to establish a symmetrical interval where eigenvalues can be located. This yields the choice $\zeta_{\mathrm{opt}} \coloneqq \frac{1}{2}\big(\big\|H_x\big\|_{\mathcal{L}(X,X^*)} + \kappa_1(x)\big)$. In particular, we thus have

$$\tau \coloneqq \omega + \zeta_{\mathrm{opt}} = \omega + \frac{1}{2}\big(\big\|H_x\big\|_{\mathcal{L}(X,X^*)} + \kappa_1(x)\big) \geq \omega + \frac{|\kappa_1(x)| + \kappa_1(x)}{2} \geq \omega + \kappa_1(x) > -\kappa_2$$

by our choice of $\omega > -\big(\kappa_1(x) + \kappa_2\big)$ and consequently obtain

$$\big\|(\tau-\omega)\mathfrak{R} - H_x\big\|_{\mathcal{L}(X,X^*)} = \big\|H_x^\tau\big\|_{\mathcal{L}(X,X)} = \big\|H_x\big\|_{\mathcal{L}(X,X^*)} - \zeta_{\mathrm{opt}} = \frac{1}{2}\big(\big\|H_x\big\|_{\mathcal{L}(X,X^*)} - \kappa_1(x)\big)\,.$$

Inserting this into the above estimate (4.1.7), we obtain (4.1.6) which completes the proof. □

Remark. *Note here that the abbreviated constant* $\mathcal{H} \coloneqq \frac{\big\|H_x\big\|_{\mathcal{L}(X,X^*)} - \kappa_1(x)}{2(\tau+\kappa_2)}$ *in* (4.1.5) *is strictly smaller than one by our choice of the "outer" regularization parameter* ω*. This can be directly retraced by the following simple computation:*

$$\begin{aligned}2(\tau+\kappa_2) &= 2(\omega+\kappa_2) + \big\|H_x\big\|_{\mathcal{L}(X,X^*)} + \kappa_1(x)\\ &> -2\kappa_1(x) + \big\|H_x\big\|_{\mathcal{L}(X,X^*)} + \kappa_1(x) = \big\|H_x\big\|_{\mathcal{L}(X,X^*)} - \kappa_1(x)\,.\end{aligned}$$

For the next result, we take advantage of the solution property of exactly computed update steps from (3.2.13).

Proposition 4.1.2: Gradient Mapping of the Regularized Model with Exactly Computed Update Steps

Let $\Delta x(\omega)$ be an exactly computed update step as in (3.2.13) at some $x \in X$. Then, for any $\tau > -\kappa_2$, the following identity holds:

$$G_\tau^{\hat{F}_{x,\omega}}\big(x + \Delta x(\omega)\big) = 0\,. \tag{4.1.8}$$

Proof. We consider the minimization problem within brackets in the definition of the regularized composite gradient mapping in (4.1.1). Here, we have to insert the derivative ϕ' of the smooth part of the regularized model $\hat{F}_{x,\omega}$ as in (4.1.4) evaluated at $y = x + \Delta x(\omega)$ which yields

$$\arg\min_{\delta x \in X} \big[f'(x) + \big(H_x + \omega\mathfrak{R}\big)\Delta x(\omega)\big]\delta x + \frac{\tau}{2}\big\|\delta x\big\|_X^2 + g\big(x + \Delta x(\omega) + \delta x\big) - g\big(x + \Delta x(\omega)\big)\,. \tag{4.1.9}$$

In the smooth case of $g = 0$, we recognize that the exactly updated iterate is characterized by the regularized Newton-system $\big(H_x + \omega\mathfrak{R}\big)\Delta x(\omega) = -f'(x)$ and consequently all terms except for the squared norm term disappear. This directly shows (4.1.8) for any $\tau \geq 0$.

The general case of $g \neq 0$, however, demands a bit more consideration. By strong convexity of the objective function for $\tau > -\kappa_2$, the above minimization problem has a unique solution $\overline{\delta x} \in X$. First order optimality conditions imply that this solution then satisfies the dual space inclusion

$$0 \in f'(x) + \big(H_x + \omega\mathfrak{R}\big)\Delta x(\omega) + \partial_F g\big(x + \Delta x(\omega) + \overline{\delta x}\big) + \tau\mathfrak{R}\overline{\delta x} \tag{4.1.10}$$

for the Fréchet-subdifferential $\partial_F g$. Note here that the exactly computed update step $\Delta x(\omega)$ as a solution of (3.2.13) suffices

$$0 \in f'(x) + \big(H_x + \omega\mathfrak{R}\big)\Delta x(\omega) + \partial_F g\big(x + \Delta x(\omega)\big)$$

which directly yields that $\delta\bar{x} = 0$ satisfies (4.1.10) and is thereby the unique solution of (4.1.9). This completes the proof of (4.1.8). □

Let us now consider the difference of gradient mappings of the objective function F and its modified second order model $\hat{F}_{x,\omega}$ at optimal solutions x_* of problem (3.2.1).

For the following we require f' to be semi-smooth near an optimal solution $x_* \in X$ of our problem (3.2.1) with respect to $H\colon X \to \mathcal{L}(X, X^*)$ as formulated in the previous section via the approximation property (3.2.11). This is where semi-smoothness enters our modified convergence theory using composite gradient mappings:

Lemma 4.1.3: Estimate for the Regularized Model in the Gradient Mapping

Let the semi-smoothness assumption (3.2.11) for f' hold at some point $x_* \in X$.

Then, the regularized composite gradient mapping satisfies the following estimate for each $\tau > -\kappa_2$ in the limit of $x \to x_*$:

$$\big\|G_\tau^F(x_*) - G_\tau^{\hat{F}_{x,\omega}}(x_*)\big\|_X \leq o\big(\big\|x_* - x\big\|_X\big) + \frac{\tau\,\omega}{\tau + \kappa_2}\big\|x_* - x\big\|_X\,.$$

Proof. The proof here follows immediately by the characterization of the regularized composite gradient mapping as in (4.1.2) and the semi-smoothness of f' according to (3.2.11). To go into detail, we start with the estimate from Proposition 3.2.3 and from there can perform the following computation:

$$\begin{aligned}
&\big\|G_\tau^F(x_*) - G_\tau^{\hat{F}_{x,\omega}}(x_*)\big\|_X \\
&\qquad = \tau\big\|\mathcal{P}_g^{\tau\mathfrak{R}}\big(\tau\mathfrak{R}x_* - f'(x_*)\big) - \mathcal{P}_g^{\tau\mathfrak{R}}\big(\tau\mathfrak{R}x_* - \big[f'(x) + \big(H_x + \omega\mathfrak{R}\big)(x_* - x)\big]\big)\big\|_X \\
&\qquad \leq \frac{\tau}{\tau+\kappa_2}\big\|\big(\tau\mathfrak{R}x_* - f'(x_*)\big) - \big(\tau\mathfrak{R}x_* - \big[f'(x) + \big(H_x + \omega\mathfrak{R}\big)(x_* - x)\big]\big)\big\|_{X^*} \\
&\qquad \leq \frac{\tau}{\tau+\kappa_2}\big\|f'(x_*) - \big(f'(x) + (H_x + \omega\mathfrak{R})(x_* - x)\big)\big\|_{X^*} \\
&\qquad = o\big(\|x_* - x\|_X\big) + \frac{\tau\,\omega}{\tau+\kappa_2}\|x_* - x\|_X\,.
\end{aligned}$$

The last identity here uses the approximation property provided by semi-smoothness of f' from (3.2.11). □

Lemma 4.1.3 already allows us to get a glimpse of the way how incorporating composite gradient mappings into our convergence theory will help us to quantify the influence of regularization on local convergence rates of our algorithm: If we manage to establish the above norm difference of gradient mappings within the local convergence proof of our (inexact) Proximal Newton method, we can see that – in addition to the common o-term arising from superlinear convergence of undamped updates – also the latter distance term, which is linear in the regularization parameter, appears.

An Existing Inexactness Criterion

While we only take advantage of composite gradient mappings in order to find useful estimates for our convergence analysis, in the literature they are predominately used in order to define inexactness criteria. We have already pointed out this circumstance when introducing inexact Proximal Newton methods from [55] in (3.1.22). Let us shortly theorize what the corresponding gradient-like inexactness criterion would look like in our scenario and discuss possible advantages and drawbacks of this formulation:

Firstly, the unregularized model which is used for step computation in the line-search approach has to be replaced with the regularized model $\hat{F}$ from (4.1.3) which we use for globalization and compensation of non-convexity. Secondly, it is sufficient to replace the upper bound on second order bilinear forms M by some constant value $\tau > \|H_x\|_{\mathcal{L}(X,X^*)} + \omega$ for all $x \in X$ and ω along the sequence of iterates generated by our method. As a consequence, at any iterate $x \in X$, the inexactness criterion from (3.1.22) in our notation generalizes to

$$\big\|G_\tau^{\hat{F}_{x,\omega}}\big(x + \Delta s(\omega)\big)\big\|_X \leq \eta\big\|G_\tau^F(x)\big\|_X \tag{4.1.11}$$

for some yet to be specified forcing term $\eta > 0$ and an inexactly computed update step candidate $\Delta s(\omega) \in X$. An obvious advantage of this formulation is its intimate relation to the finite dimensional counterpart which thus also gives an intuition concerning the proof of corresponding convergence results. Additionally, the quality of the candidate is directly measured within the gradient mapping which is – as we have seen right after its definition – a viable measure for the optimality of the current iterate. Thus, the decrease in norm

of the gradient mapping, which is also indicated by (4.1.11), already gives a hint towards globalization of the inexact method. Even though this is not considered in [55], one can easily see that additionally demanding a trivial decrease condition like

$$\hat{F}_{x,\omega}\big(x+\Delta s(\omega)\big) < \hat{F}_{x,\omega}(x) = F(x) \quad \Leftrightarrow \quad \lambda_{x,\omega}\big(\Delta s(\omega)\big) < 0 \tag{4.1.12}$$

already enables the proof of similar global convergence results as we have verified them in the exact case in Section 3.2.3. In particular, assuming that (4.1.11) holds along the sequence of iterates, one can easily prove

$$\big\|G_\tau^F(x)\big\|_X \leq \frac{\tau}{1-\eta}\big(1+\frac{\tau-\omega-\kappa_1(x)}{\tau+\kappa_2}\big)\big\|\Delta s(\omega)\big\|_X$$

and thereby the convergence of the gradient mapping to zero which then again implies global convergence via first order optimality conditions of the corresponding computational problem. Also the transition to local convergence as formulated in Section 3.2.5 turns out to be achievable under these requirements.

So what are the drawbacks of this approach to inexactness if the convergence results derivable from it look so convincing? Again, as has been the motivation for the alternative sufficient decrease criterion addressed in Section 3.2.6, the occurring problems are rather of computational than of theoretical nature. The crucial peculiarity again hides in the adequate structural representation of function spaces via a multi-level approximation rather than a diagonal one as would be fitting for the finite dimensional case.

The latter structural assumption enables an easy and efficient evaluation of the gradient mapping (and thus also Proximal Gradient steps) due to the diagonal structure of the norm term while in an infinite dimensional setting the computation of the gradient mapping is for this reason quite demanding, even similarly expensive as computing the actual exact update step $\Delta x(\omega)$ from (3.2.13). Consequently, evaluating (4.1.11) for every iteration within the subproblem solver quickly becomes very costly and thereby immediately eclipses the savings which we could potentially gain from inexactly computing the update steps. For this reason, we will resort to a different inexactness criterion.

4.1.2 First Inexactness Criterion and Local Convergence

As pointed out beforehand, we do not use an inexactness criterion of the form (4.1.11) due to its immense computational effort in function space. Instead, we exploit the advantageous properties of the TNNMG subproblem solver by resorting to an actual relative error estimate of the form

$$\big\|\Delta x(\omega)-\Delta s(\omega)\big\|_X \leq \eta\big\|\Delta x(\omega)\big\|_X \tag{4.1.13}$$

where $\Delta x(\omega)$ denotes the exact solution of the update step computation subproblem (3.2.13) and $\Delta s(\omega)$ is the corresponding inexact candidate. The influence of the forcing terms $\eta \geq 0$ on local convergence rates will be investigated in Theorem 4.1.4.

Before actually stating the local convergence results, let us remark that the inexactness criterion (4.1.13) is trivially satisfied by exactly computed update steps and the forcing term η can be understood as a measure for the margin for error which we allow in the computation. Additionally, the fact that the inexactly computed update steps $\Delta s(\omega)$ are in our case iterates from the linearly convergent TNNMG subproblem solver implies that – sooner or later within the solution process of (3.2.13) – the requirement (4.1.13) will be satisfied.

Furthermore, let us comment on the efficient evaluation of this relative error estimate. At first sight, this is not completely obvious since apparently we do not have the exact solution $\Delta x(\omega)$ of the update computation subproblem (3.2.13) at hand. In order to deal with this issue, we take advantage of the multigrid structure of the iterative subproblem solver which we employ, i.e., the TNNMG method from [33]. By δ^j we denote TNNMG corrections, let therefore $\Delta s^i(\omega) = \sum_{j=1}^{i} \delta^j$ be an iterate within the inner solver towards the exact solution $\Delta x(\omega)$ and θ the "constant" multigrid convergence rate from $\|\delta^j\|_X \leq \theta \|\delta^{j-1}\|_X$. A simple computation provides us with

$$\|\Delta x(\omega) - \Delta s^i(\omega)\|_X = \sum_{j=i+1}^{\infty} \|\delta^j\|_X \leq \|\delta^i\|_X \sum_{j=i+1}^{\infty} \theta^{j-i} = \frac{\theta}{1-\theta} \|\delta^i\|_X .$$

Similarly, for the norm of the exact solution we obtain

$$\begin{aligned} \|\Delta x(\omega)\|_X &= \|\sum_{j=1}^{\infty} \delta^j\|_X = \|\Delta s^i(\omega) + \sum_{j=i+1}^{\infty} \delta^j\|_X \geq \|\Delta s^i(\omega)\|_X - \|\sum_{j=i+1}^{\infty} \delta^j\|_X \\ &\geq \|\Delta s^i(\omega)\|_X - \frac{\theta}{1-\theta} \|\delta^i\|_X \end{aligned}$$

by a simple triangle inequality. Combining both of these estimates allows us to establish

$$\frac{\|\Delta x(\omega) - \Delta s^i(\omega)\|_X}{\|\Delta x(\omega)\|_X} \leq \frac{\frac{\theta}{1-\theta}\|\delta^i\|_X}{\|\Delta s^i(\omega)\|_X - \frac{\theta}{1-\theta}\|\delta^i\|_X} \overset{!}{\leq} \eta \tag{4.1.14}$$

as a sufficient and easy to evaluate alternative inexactness criterion for the relative error estimate (4.1.13). Numerical experiments, which we also incorporated to Section 4.1.6, clearly demonstrate that the performed triangle inequalities are sharper than one might have expected. Thus, the evaluation of the alternative criterion from (4.1.14) comes very close to using the actual relative error for our computations later on.

Similarly using the advantageous convergence properties of the subproblem solver, it is also possible to consider a so-called *model-based approach* for the design of an inexactness criterion. There, not the relative error in norm but in the objective functional of the subproblem, i.e., the regularized second order decrease model $\lambda_{x,\omega}$ is used as an indicator for sufficient accuracy in computation. This approach features some favorable properties in particular concerning its conceptual design but problems with the corresponding convergence analysis under the assumptions stated in our framework arise. In particular, a quadratic upper bound on g similar to the lower one from (A4) is necessary in order to achieve the corresponding local convergence results which would implicitely impose a continuity assumption on our non-smooth part. This framework appears to be too restrictive for our purposes.

With the relative error inexactness criterion (4.1.13) as well as the auxiliary results concerning regularized composite gradient mappings from Section 4.1.1 and norm estimates from Lemma 3.2.26 at hand, we can now tackle the proof of the following local acceleration result:

Theorem 4.1.4: Local Convergence of the Inexact Proximal Newton Method Using the Relative Error Criterion (4.1.13)

Suppose that the semi-smoothness assumption (3.2.11) holds at an optimal solution $x_* \in X$ of (3.2.1) together with (A3) and (A4) for $\kappa_1(x) + \kappa_2 > 0$ in a neighborhood of

that x_*.

Then, the inexact Proximal Newton method with update steps $\Delta s_k(\omega_k)$ computed according to (3.2.13) at $x_k \in X$ close to x_* with the inexactness criterion (4.1.13) for $\eta_k \geq 0$ exhibits the following local convergence behavior:

(i) The sequence of iterates locally converges linearly if ω_k and η_k are sufficiently small, more precisely if there exists some constant $0 < \Theta < 1$ and $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$ the following estimate holds:

$$\frac{\big(\omega_k + \big\|H_{x_k}\big\|_{\mathcal{L}(X,X^*)} + \kappa_2\big)\eta_k + \omega_k}{\omega_k + \kappa_1(x_k) + \kappa_2} < \Theta\,. \tag{4.1.15}$$

(ii) The sequence of iterates locally converges superlinearly in case both sequences $(\omega_k)_{k\in\mathbb{N}}$ and $(\eta_k)_{k\in\mathbb{N}}$ converge to zero for $k \to \infty$.

Proof. For the sake of simplicity, we will omit the sequence indices of all quantities here and denote by $x = x_k$, $\omega = \omega_k$ and $\eta = \eta_k$ the current iterate, regularization parameter and forcing term. For the next iterate, we write $x_+(\omega) = x + \Delta s(\omega) = x_k + \Delta s_k(\omega_k)$ and $H_x = H_{x_k}$ stands for the current second order bilinear form.

Additionally, we fix $\tau \coloneqq \omega + \frac{1}{2}\big(\big\|H_x\big\|_{\mathcal{L}(X,X^*)} + \kappa_1(x)\big)$ for the gradient mapping regularization parameter which allows us to take advantage of the auxiliary estimates deduced in Lemma 4.1.1. Under these circumstances, the first part of (4.1.5) from Lemma 4.1.1 provides us with

$$\begin{aligned}\big\|x_+(\omega) - x_*\big\|_X &\leq \frac{1}{\tau\big(1-\mathcal{H}\big)}\big\|G_\tau^{\hat{F}_{x,\omega}}(x + \Delta s(\omega)) - G_\tau^{\hat{F}_{x,\omega}}(x_*)\big\|_X \\ &\leq \frac{1}{\tau\big(1-\mathcal{H}\big)}\Big[\big\|G_\tau^{\hat{F}_{x,\omega}}(x + \Delta s(\omega))\big\|_X + \big\|G_\tau^{\hat{F}_{x,\omega}}(x_*)\big\|_X\Big]\end{aligned} \tag{4.1.16}$$

where we have accordingly abbreviated the constant $\mathcal{H} \coloneqq \frac{\big\|H_x\big\|_{\mathcal{L}(X,X^*)} - \kappa_1(x)}{2(\tau+\kappa_2)} < 1$. As a next step, we take a look at the first norm term in brackets in (4.1.16). We use (4.1.8) from Proposition 4.1.2 together with the second part of (4.1.5) from Lemma 4.1.1 for $y \coloneqq x + \Delta s(\omega)$ and $z \coloneqq x + \Delta x(\omega)$ in order to obtain the following estimate:

$$\begin{aligned}\big\|G_\tau^{\hat{F}_{x,\omega}}\big(x + \Delta s(\omega)\big)\big\|_X &= \big\|G_\tau^{\hat{F}_{x,\omega}}\big(x + \Delta s(\omega)\big) - G_\tau^{\hat{F}_{x,\omega}}\big(x + \Delta x(\omega)\big)\big\|_X \\ &\leq \tau\big(1+\mathcal{H}\big)\big\|\Delta x(\omega) - \Delta s(\omega)\big\|_X\,.\end{aligned}$$

For the ensuing norm difference, we take advantage of the relative error inexactness criterion (4.1.13) together with the monotonicity of update step norms concerning the damping parameter ω as in Lemma 3.2.26. Additionally, the superlinear convergence for full and exactly computed update steps close to optimal solutions proven in Theorem 3.2.8 is important here:

$$\big\|\Delta x(\omega) - \Delta s(\omega)\big\|_X \leq \eta\big\|\Delta x(\omega)\big\|_X \leq \eta\big\|\Delta x\big\|_X \leq o\big(\big\|x - x_*\big\|_X\big) + \eta\big\|x - x_*\big\|_X\,. \tag{4.1.17}$$

By the stationarity of x_* together with Lemma 4.1.3, for the second term in brackets in (4.1.16) we have

$$\big\|G_\tau^{\hat{F}_{x,\omega}}(x_*)\big\|_X = \big\|G_\tau^{\hat{F}_{x,\omega}}(x_*) - G_\tau^{F}(x_*)\big\|_X \leq o\big(\big\|x - x_*\big\|_X\big) + \frac{\omega\tau}{\tau+\kappa_2}\big\|x - x_*\big\|_X\,. \tag{4.1.18}$$

The estimates (4.1.17) and (4.1.18) suffice to quantify the influence of either inexactness or damping on local convergence rates of our algorithm. Inserting both of them into (4.1.16) above yields

$$\left\|x_+(\omega) - x_*\right\|_X \leq \frac{(1+\mathcal{H})\eta + \frac{\omega}{\tau+\kappa_2}}{1-\mathcal{H}}\left\|x - x_*\right\|_X + o\big(\left\|x - x_*\right\|_X\big)\,. \tag{4.1.19}$$

All that remains to do now is simplify the rather complicated prefactor term within the estimate above. We expand the fraction by $2(\tau+\kappa_2)$ and use that by the definition of τ we have

$$2(\tau+\kappa_2) = 2(\omega+\kappa_2) + \left\|H_x\right\|_{\mathcal{L}(X,X^*)} + \kappa_1(x)\,.$$

This provides us with

$$\frac{(1+\mathcal{H})\eta + \frac{\omega}{\tau+\kappa_2}}{1-\mathcal{H}} = \frac{\big(2(\tau+\kappa_2) + \left\|H_x\right\|_{\mathcal{L}(X,X^*)} - \kappa_1(x)\big)\eta + 2\omega}{2(\tau+\kappa_2) - \left\|H_x\right\|_{\mathcal{L}(X,X^*)} + \kappa_1(x)} = \frac{\big(\omega + \left\|H_x\right\|_{\mathcal{L}(X,X^*)} + \kappa_2\big)\eta + \omega}{\omega + \kappa_1(x) + \kappa_2}$$

Inserting this identity into (4.1.19) now directly yields

$$\left\|x_+(\omega) - x_*\right\|_X \leq \frac{\big(\omega + \left\|H_x\right\|_{\mathcal{L}(X,X^*)} + \kappa_2\big)\eta + \omega}{\omega + \kappa_1(x) + \kappa_2}\left\|x - x_*\right\|_X + o\big(\left\|x - x_*\right\|_X\big)\,. \tag{4.1.20}$$

From here, both of the asserted cases for local convergence behavior are an immediate consequence of (4.1.20). □

Remark. *The estimate* (4.1.15) *yields a couple of algorithmically relevant insights. Firstly, the linear convergence factor* Θ *can only be small if both* ω_k *and* η_k *are small. Hence, computing steps very accurately does only pay off if* ω_k *is very small. We will see in Section 4.1.4 that close to optimal solutions arbitrarily small regularization parameters* $\omega_k \approx 0$ *can indeed be used also in the currently considered inexact scenario.*

Secondly, if we neglect $\omega_k \approx 0$ *in these later stages of the minimization process,* (4.1.15) *simplifies to*

$$\frac{\left\|H_{x_k}\right\|_{\mathcal{L}(X,X^*)} + \kappa_2}{\kappa_1(x) + \kappa_2}\eta_k \leq \Theta,$$

where the prefactor on the left hand side can be interpreted as a local condition number of the composite problem. Indeed, for $\kappa_2 = 0$ *(, i.e., the smooth or convex case,) it coincides with the condition number of* H_x *relative to* $\|\cdot\|_X$*. Thus, to achieve a given rate of local convergence,* η_k *has to be chosen tighter the higher the condition number. This additionally underlines the necessity of an adequate choice of function space* X *and norm* $\|\cdot\|_X$*.*

As we can see, the local superlinear convergence as deduced in Theorem 3.2.8 remains true also in the inexact case if we reduce the value of the forcing terms to zero as we approach optimal solutions of our composite minimization problem. Apparently, this also means that close to optimal solutions the margin for error which we accept for an inexact update step has to tend to zero which is apparent since we want to identify optimal solutions also with high accuracy.

Additionally, we have been able to extend the local convergence result from Theorem 3.2.8 insofar that we quantified the influence of damping update steps on (local) convergence rates. We are now also aware of more insightful criteria both for linear and superlinear convergence of our method. This helps us understand the process of local convergence of the (inexact) Proximal Newton method to an even greater extent.

4.1.3 Second Inexactness Criterion and Global Convergence

Now that we have clarified the local convergence properties of our inexact Proximal Newton method depending on the forcing terms in criterion (4.1.13), we will take into consideration whether the globalization scheme via the additional norm term in (3.2.13) still fulfills its purpose and yields some global convergence results.

Cauchy Decrease Steps and the Subgradient Model

As we have already pointed out when describing the predominant gradient-like inexactness criterion (4.1.11), for this formulation it suffices to demand the trivial decrease condition (4.1.12) in order to achieve global convergence results of the ensuing methods. Since we have chosen not to use this approach to inexactness due to structural peculiarities of the function space framework, we have to put some more deliberations into the globalization of our inexact method.

To this end, we will now introduce a second crucial criterion which the inexactly computed update steps $\Delta s_k(\omega_k)$ have to satisfy in order to be admissible for our method. It can be viewed as an adopted strategy from smooth trust region methods where rather cheap so-called *Cauchy decrease steps* are used to measure functional value descent for the actual update steps, cf. e.g. [17, Chapter 6.3].

There are several conceivable ways to define and compute such comparative Cauchy decrease steps. A canonical choice would be a simple Proximal Gradient step, i.e., the minimizer of the regularized linear model $\lambda^C_{x,\hat{\omega}} \colon X \to]-\infty, \infty]$ defined by

$$\lambda^C_{x,\hat{\omega}}(\delta x) \coloneqq f'(x)\delta x + \frac{\hat{\omega}}{2}\big\|\delta x\big\|_X^2 + g(x+\delta x) - g(x)$$

for some regularization parameter $\hat{\omega} \geq 0$. As was the problem with evaluating the gradient mapping for our first inexactness criterion, also the minimization of $\lambda^C_{x,\hat{\omega}}$ is similarly expensive as computing the exact Proximal Newton step right away in our general Hilbert space setting. Thus, the idea arises to find some comparative update step which we can compute with marginal effort once per "outer iterations" in order to measure its functional value descent and then compare it to our inexact update step candidates.

To this end, we define the *regularized subgradient decrease model* of F around $x \in X$ with respect to $\mu \in \partial_F g(x)$ and a regularization parameter $\hat{\omega} > 0$ by

$$\lambda^\mu_{x,\hat{\omega}} \colon X \to \mathbb{R}\,, \quad \lambda^\mu_{x,\hat{\omega}}(\delta x) \coloneqq f'(x)\delta x + \mu\,\delta x + \frac{\hat{\omega}}{2}\big\|\delta x\big\|_X^2 \tag{4.1.21}$$

and we refer to the respective minimizer

$$\Delta x^\mu(\hat{\omega}) \coloneqq \operatorname*{arg\,min}_{\delta x \in X} \lambda^\mu_{x,\hat{\omega}}(\delta x) \tag{4.1.22}$$

as the corresponding *subgradient step*. Before introducing the second inexactness criterion which makes use of the above model and step, we will establish an analytical connection between (4.1.21) and our initially defined regularized second order decrease model $\lambda_{x,\omega}$ from (4.1.3). To this end, we remember that the regularization parameter $\omega \geq 0$ is generally chosen such that the modified non-smooth part

$$g^{H_x}_\omega \colon X \to]-\infty, \infty] \quad , \quad g^{H_x}_\omega(x) \coloneqq g(x) + \frac{1}{2}\big(H_x + \omega\mathfrak{R}\big)(x)^2$$

is convex and thus the subproblem (3.2.13) allows for a unique solution. Consequently, the characterization of the convex subdifferential $\partial\tilde{g}(x)$ yields that for any $\tilde{\mu} = \mu + (H_x + \omega\mathfrak{R})x \in \partial g_\omega^{H_x}(x)$ with $\mu \in \partial_F g(x)$ we have that

$$g_\omega^{H_x}(x+\delta x) \geq g_\omega^{H_x}(x) + \tilde{\mu}\,\delta x \quad \text{ and thus } \quad g(x+\delta x) - g(x) + \frac{1}{2}H_x(\delta x)^2 + \frac{\omega}{2}\big\|\delta x\big\|_X^2 \geq \mu\,\delta x$$

holds for any $\delta x \in X$ and $\mu \in \partial_F g(x)$. We immediately obtain that

$$\begin{aligned}\lambda_{x,\hat{\omega}}^{\mu}(\delta x) &= f'(x)\delta x + \frac{\hat{\omega}}{2}\big\|\delta x\big\|_X^2 + \mu\delta x \\ &\leq f'(x)\delta x + \frac{1}{2}H_x(\delta x)^2 + \frac{\hat{\omega}+\omega}{2}\big\|\delta x\big\|_X^2 + g(x+\delta x) - g(x) = \lambda_{x,\hat{\omega}+\omega}(\delta x)\end{aligned} \tag{4.1.23}$$

is true for any $\delta x \in X$. In particular, this estimate apparently also holds for the respective minima of the decrease models of the composite objective function. For that reason, from (4.1.23) we obtain

$$\lambda_{x,\hat{\omega}}^{\mu}\big(\Delta x^{\mu}(\hat{\omega})\big) \leq \lambda_{x,\hat{\omega}+\omega}\big(\Delta x(\hat{\omega}+\omega)\big) \leq -\frac{1}{2}\big(\hat{\omega}+\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\hat{\omega}+\omega)\big\|_X^2 \tag{4.1.24}$$

for any $\hat{\omega} > 0$ where the last estimate constitutes a result from the exact case in (3.2.16) and will give us norm-like descent in the objective functional later on. Obviously, we now want to link this norm-like decrease within the subgradient model to the regularized second order decrease model $\lambda_{x,\omega}\big(\Delta s(\omega)\big)$ for our inexactly computed update step $\Delta s(\omega)$ and lastly to the direct descent within the objective functional F.

Second Inexactness Criterion and Efficient Evaluation

We will establish the first one of these connections via the actual second inexactness criterion which will thus also be checked within our algorithm and implementation. For this purpose, it is sufficient if an inexactly computed update step $\Delta s(\omega)$ satisfies the estimate

$$\lambda_{x,\omega}\big(\Delta s(\omega)\big) \leq \lambda_{x,\tilde{\omega}}^{\mu}\big(\Delta x^{\mu}(\tilde{\omega})\big) \quad \text{for some} \quad \tilde{\omega} < \tilde{\omega}_{\max} \tag{4.1.25}$$

where the upper bound $\tilde{\omega}_{\max} > 0$ is the *subgradient regularization bound* the value of which is yet to be specified. This inequality now constitutes our formal second inexactness criterion which we will also refer to as the *subgradient inexactness criterion.*

Let us shortly elaborate on the efficient evaluation of this estimate and from there derive the actual implementation of the criterion: The solution property of $\Delta x^{\mu}(\tilde{\omega})$ provides us with first order conditions for the corresponding minimization problem in the form of

$$0 = f'(x) + \mu + \hat{\omega}\mathfrak{R}\Delta x^{\mu}(\tilde{\omega})$$

and thus $\Delta x^{\mu}(\tilde{\omega}) = -(\tilde{\omega}\mathfrak{R})^{-1}\big(f'(x)+\mu\big)$. For a given value of $\lambda_{x,\omega}\big(\Delta s(\omega)\big)$, i.e., decrease along an inexactly computed update step within the regularized second order model, we can thus theoretically determine $\tilde{\omega}$ such that (4.1.25) is satisfied with equality. This can be seen as follows:

$$\begin{aligned}\lambda_{x,\omega}\big(\Delta s(\omega)\big) &\overset{!}{=} \lambda_{x,\tilde{\omega}}^{\mu}\big(\Delta x^{\mu}(\tilde{\omega})\big) = \big(f'(x)+\mu\big)\Delta x^{\mu}(\tilde{\omega}) + \frac{\tilde{\omega}}{2}\big\|\Delta x^{\mu}(\tilde{\omega})\big\|_X^2 \\ &= \big(f'(x)+\mu\big)\big[-(\tilde{\omega}\mathfrak{R})^{-1}\big(f'(x)+\mu\big)\big] + \frac{\tilde{\omega}}{2}\big\|-(\tilde{\omega}\mathfrak{R})^{-1}\big(f'(x)+\mu\big)\big\|_X^2 \\ &= -\frac{1}{2\tilde{\omega}}\big\|f'(x)+\mu\big\|_{X^*}^2 .\end{aligned} \tag{4.1.26}$$

The above computation provides us with the theoretical value

$$\tilde{\omega} = -\frac{\left\|f'(x)+\mu\right\|_{X^*}^2}{2\lambda_{x,\omega}\big(\Delta s(\omega)\big)} \overset{!}{<} \tilde{\omega}_{\max} \tag{4.1.27}$$

for the regularization parameter within the subgradient minimization problem (4.1.22). This quantity should remain bounded in order to enable the proof of global convergence results later on. Thus, as also pointed out in (4.1.27), we have established a sufficient estimate for our subgradient inexactness criterion (4.1.25) by demanding boundedness of $\tilde{\omega}$ from above by $\tilde{\omega}_{\max}$. Note here that – as can be seen in (4.1.26) – the value for $\lambda^{\mu}_{x,\tilde{\omega}}\big(\Delta x^{\mu}(\tilde{\omega})\big)$ increases as $\tilde{\omega}$ does.

Since globalization mechanisms in general should only provide worst case estimates and not slow down the convergence of our algorithm, we want the subgradient inexactness criterion to only interfere with update step computation on rare occasions and thus choose $\tilde{\omega}_{\max}$ very large.

The dual norm occurring in the numerator of (4.1.27) can be evaluated as follows: We compute the minimizer of the subgradient model $\Delta x^{\mu}(1) \in X$ from (4.1.22) and afterwards evaluate the linear functional $f'(x)+\mu \in X^*$ there. Here, the Fréchet-subdifferential element $\mu \in \partial_F g(x)$ is chosen such that the norm $\left\|f'(x)+\mu\right\|_{X^*}$ is minimal. Obviously, this depends on the specific minimization problem at hand but due to the non-smooth nature of g, it is often possible to exploit the set-valued subdifferential for this purpose.

Satisfiability and Algorithmic Strategy of the Subgradient Criterion

Let us add some remarks concerning satisfiability of the subgradient inexactness criterion: As mentioned above, the freedom of choice of μ within $\partial_F g(x)$ opens up possibilities to decrease the value of $\left\|f'(x)+\mu\right\|_{X^*}$ right away. Additionally, considering the exact case for update step computation is very insightful in order to see that the criterion will be fulfilled by late iterations of the inner solver. For establishing plausibility of that circumstance, it is useful to interpret $\left\|f'(x)+\mu\right\|_{X^*} \approx \operatorname{dist}\big(\partial_F F(x), 0\big)$, i.e., to assume that $\mu \in \partial_F g(x)$ is chosen (nearly) optimally for our purpose of finding solutions of (3.2.1).

Then, we can take a look back at the global convergence arguments from the exact case in Theorem 3.2.14 and recognize that the abovementioned dual space norm scales with the squared update step norm over the course of our algorithm. Together with the fact that the denominator term, i.e., the modified second order decrease model $\lambda_{x,\omega}\big(\Delta x(\omega)\big)$, exhibits the same behavior as we have shown in (3.2.16), this suggests that the subgradient inexactness criterion (4.1.25) is eventually fulfilled by iterates of convergent subproblem solvers if the corresponding constant $\tilde{\omega}_{\max}$ is chosen sufficiently large.

For this reason, we use a very large $\tilde{\omega}_{\max}$ within specific implementations of our algorithm. Then, we can also be sure that our globalization techniques notably interfere with the minimization process only on rare exceptional occasions. Also note that our notion of "convergent subproblem solvers" in this context particularly incorporates the convergence of objective values. Generally, by convergence of the solver only convergence of the corresponding iterates is ensured. Convergence of objective values in addition might require some continuity assumptions on the non-smooth part g which we generally do not have here.

The algorithmic strategy behind the subgradient inexactness criterion can now be summarized as follows: For the present iterate of the outer loop $x \in X$, we solve the linearized

problem (4.1.22) for the computation of the dual norm $\big\|f'(x)+\mu\big\|_{X^*}$ and initiate the inner loop in order to determine the next inexact update step. At every iterate $\Delta s(\omega)$ of the inner solver for subproblem (3.2.13), we compute the corresponding subgradient regularization parameter $\tilde{\omega}$ from (4.1.27) and check $\tilde{\omega} < \tilde{\omega}_{\max}$. As a consequence of our satisfiability discussion, either $\tilde{\omega}_{\max}$ is chosen large enough and we will eventually achieve $\tilde{\omega} < \tilde{\omega}_{\max}$ for some inexact step or we will compute an exact update step $\Delta x(\omega)$ which on its own provides us with global convergence of the sequence of iterates as presented for the exact scenario in Section 3.2.3.

Summary of Inexactness Criteria

With both of our inexactness criteria at hand, let us shortly reflect on their computational effort and compare them to possible alternatives: For the relative error criterion (4.1.13) in its form (4.1.14) only the evaluation of the fraction and its comparison to the forcing term is necessary since all occurring norms are already present within the subproblem solver. The subgradient inexactness criterion as described before requires the solution of the quadratic minimization problem (4.1.22) once per outer iteration of our method together with the evaluation of the quadratic model $\lambda_{x,\omega}\big(\Delta s(\omega)\big)$ at each inner iteration which is a cheap operation.

For comparative algorithms from literature, cf. [11, 44, 55], the gradient-like inexactness criterion (4.1.11) has to be assessed at every inner iteration together with one comparison of the second order decrease model value with its base value for $\delta x = 0$. As mentioned before, the former operation is very costly for non-diagonal function space norm representations, particularly in comparison to solving a linearized problem once per outer iteration. This emphasizes both the necessity and the benefit of our adjustments to existing inexactness criteria. The summarized procedure can be retraced in the scheme of Algorithm 11.

Sufficient Decrease Criterion and Global Convergence

For global convergence in the case of inexactly computed update steps with the criteria introduced above, we still have to carry out some more deliberations. The last missing ingredient in our recipe for norm-like descent within the composite objective functional is an adequate sufficient decrease criterion. For that reason, we will reuse the corresponding formulation from the exact case in (3.2.15). We continue to use the concept from Definition 3.2.11 and thus say that an inexactly computed update step $\Delta s(\omega)$ is admissible for sufficient decrease if for some prescribed $\gamma \in]0,1]$ the estimate

$$F\big(x+\Delta s(\omega)\big) - F(x) \leq \gamma \lambda_{x,\omega}\big(\Delta s(\omega)\big) \tag{4.1.28}$$

holds. Also here, we have to ensure that adequately defined strategies for finding regularization parameters are finite, i.e., we have to justify that (4.1.28) holds for sufficiently large values of the regularization parameter ω.

Before we do so, let us get a prospect of the results we can achieve once (4.1.28) is satisfied. To this end, we combine estimates (4.1.28), (4.1.25), the monotonicity of $\lambda^{\mu}_{x,\tilde{\omega}}\big(\Delta x^{\mu}(\tilde{\omega})\big)$ with

respect to $\tilde{\omega}$ as well as (4.1.24) from above and thus recognize that we obtain

$$\begin{aligned} F\big(x+\Delta s(\omega)\big) - F(x) &\leq \gamma\lambda_{x,\omega}\big(\Delta s(\omega)\big) = \gamma\lambda^{\mu}_{x,\tilde{\omega}}\big(\Delta x^{\mu}(\tilde{\omega})\big) \leq \gamma\lambda^{\mu}_{x,\tilde{\omega}+1}\big(\Delta x^{\mu}(\tilde{\omega}+1)\big) \\ &\leq -\frac{(\tilde{\omega}+\omega+1+\kappa_1(x)+\kappa_2)\gamma}{2}\big\|\Delta x(\tilde{\omega}+\omega+1)\big\|_X^2 \\ &\leq -\frac{\gamma}{2}\big\|\Delta x(\tilde{\omega}_{\max}+\omega+1)\big\|_X^2 . \end{aligned} \tag{4.1.29}$$

Note that we additionally used $\tilde{\omega} \geq 0$ and $\omega+\kappa_1(x)+\kappa_2 \geq 0$ as well as $\tilde{\omega} < \tilde{\omega}_{\max}$ together with the monotonicity result from Lemma 3.2.26. We can see that once an inexact update step yields sufficient decrease according to (4.1.28), we again have norm-like decrease in the composite objective functional for some (stronger regularized) exactly computed update step which will then again provide us with first order optimality conditions leading to global convergence results. This achievement can be contributed to the subgradient inexactness criterion (4.1.25) since this estimate allowed us to establish an inequality between our inexact candidate for the update and an exactly computed counterpart which we can take advantage of in convergence analysis.

Before the formulation of the ensuing global convergence results, we have to consider the following lemma which ensures finiteness of adequately defined backtracking procedures, i.e., that (4.1.28) is satisfied as soon as ω is large enough.

Lemma 4.1.5: Satisfiability of the Sufficient Decrease Criterion (4.1.28) in the Inexact Case

The sufficient decrease criterion (4.1.28) is fulfilled by inexactly computed update steps $\Delta s(\omega)$ which additionally satisfy the inexactness criteria (4.1.13) and (4.1.25) if the regularization parameter ω satisfies the inequality

$$\frac{1-\gamma}{(1+\eta)^2}(\omega+\kappa)^2 + \omega(\omega+\tilde{\omega}_{\max}+\kappa-L) \geq L(\tilde{\omega}_{\max}+\kappa)$$

where we have abbreviated $\kappa \coloneqq \kappa_1(x)+\kappa_2$ and $L \coloneqq L_f - \kappa_1(x)$.

Proof. The first inexactness criterion (4.1.13) provides us with the norm estimate

$$\big\|\Delta s(\omega)\big\|_X \leq \big\|\Delta s(\omega)-\Delta x(\omega)\big\|_X + \big\|\Delta x(\omega)\big\|_X \leq (1+\eta)\big\|\Delta x(\omega)\big\|_X . \tag{4.1.30}$$

Additionally, with the aid of the second inexactness criterion (4.1.25), the estimate of the subgradient model against the exact update step norm (4.1.24), and the equivalence result in update step norms from Lemma 3.2.26 we obtain

$$\begin{aligned} \lambda_{x,\omega}\big(\Delta s(\omega)\big) = \lambda^{\mu}_{x,\tilde{\omega}}\big(\Delta x^{\mu}(\tilde{\omega})\big) &\leq \lambda^{\mu}_{x,\tilde{\omega}_{\max}}\big(\Delta x^{\mu}(\tilde{\omega}_{\max})\big) \\ &\leq -\frac{1}{2}\big(\tilde{\omega}_{\max}+\omega+\kappa_1(x)+\kappa_2\big)\big\|\Delta x(\tilde{\omega}_{\max}+\omega)\big\|_X^2 \\ &\leq -\frac{\big(\omega+\kappa_1(x)+\kappa_2\big)^2}{2\big(\tilde{\omega}_{\max}+\omega+\kappa_1(x)+\kappa_2\big)}\big\|\Delta x(\omega)\big\|_X^2 . \end{aligned}$$

Combining this estimate with the one from (4.1.30) now provides us with

$$\lambda_{x,\omega}\big(\Delta s(\omega)\big) \leq -\frac{\big(\omega+\kappa_1(x)+\kappa_2\big)^2}{2(1+\eta)^2\big(\tilde{\omega}_{\max}+\omega+\kappa_1(x)+\kappa_2\big)}\big\|\Delta s(\omega)\big\|_X^2\,. \tag{4.1.31}$$

Here, we recognize that the inequality from the assertion is equivalent to

$$\frac{L_f-\kappa_1-\omega}{2}\cdot\frac{2\big(\tilde{\omega}_{\max}+\omega+\kappa_1(x)+\kappa_2\big)(1+\eta)^2}{\big(\omega+\kappa_1(x)+\kappa_2\big)^2}\leq 1-\gamma$$

which together with (4.1.31) and $x_+(\omega)=x+\Delta s(\omega)$ lets us infer by the Lipschitz continuity of f' with constant L_f that

$$\begin{aligned}F\big(x_+(\omega)\big)-F(x)&\leq f'(x)\Delta s(\omega)+\frac{L_f}{2}\big\|\Delta s(\omega)\big\|_X^2+g\big(x_+(\omega)\big)-g(x)\\&\leq\lambda_{x,\omega}\big(\Delta s(\omega)\big)+\frac{1}{2}(L_f-\kappa_1(x)-\omega)\big\|\Delta s(\omega)\big\|_X^2\\&\leq\lambda_{x,\omega}\big(\Delta s(\omega)\big)-(1-\gamma)\lambda_{x,\omega}\big(\Delta s(\omega)\big)=\gamma\lambda_{x,\omega}\big(\Delta s(\omega)\big)\end{aligned}$$

holds and we conclude that $\Delta s(\omega)$ yields sufficient decrease according to (4.1.28). □

Remark. *As we have done after formulating the corresponding result for the exact case in Lemma 3.2.12, we can also here add some deliberations which simplify the rather complicated requirement on the regularization parameter from above in qualified situations. Firstly, we note that the sum*

$$\omega+\kappa=\omega+\kappa_1(x)+\kappa_2>0$$

is generally perceived to be very small in the sense that in this combination ω is chosen just to achieve strong convexity of the update step computation subproblem (3.2.13)*. In particular, the bound on ω from Lemma 4.1.5 is also satisfied if we neglect these terms and thus obtain*

$$\omega(\tilde{\omega}_{\max}-L_f)\geq L_f\big(\tilde{\omega}_{\max}+\kappa_1(x)+\kappa_2\big)\quad\Leftrightarrow\quad\omega\geq\frac{L_f\big(\tilde{\omega}_{\max}+\kappa_1(x)+\kappa_2\big)}{\tilde{\omega}_{\max}-L_f}\,. \tag{4.1.32}$$

This bound in particular only makes sense for $\tilde{\omega}_{\max}>L_f$ which is what we assume to hold anyways since $\tilde{\omega}_{\max}$ is an algorithmic parameter which we choose very large in applications.

From another perspective, we can take a closer look at the proof and as in the exact case of Lemma 3.2.12 and perceive that due to the non-positivity of $\lambda_{x,\omega}\big(\Delta s(\omega)\big)$ from (4.1.31) *and $\gamma<1$ the bound $\omega>L_f-\kappa_1(x)$ is sufficient for establishing the admissibility of the regularization parameter and corresponding inexactly computed update step. As a consequence, also here the question arises in which cases this simple alternative bound on ω prevails in significance at least in contrast to* (4.1.32)*. In that regard, a short computation results in*

$$L_f-\kappa_1(x)<\frac{L_f\big(\tilde{\omega}_{\max}+\kappa_1(x)+\kappa_2\big)}{\tilde{\omega}_{\max}-L_f}\quad\Leftrightarrow\quad L_f\kappa_2+L_f^2+\kappa_1(x)\tilde{\omega}_{\max}>0\,.$$

Interpreting this sufficient condition, the large choice of $\tilde{\omega}_{\max}>0$ suggests that the alternative bound $\omega>L_f-\kappa_1(x)$ is more meaningful in case $\kappa_1(x)$ is positive at the current iterate. In the non-elliptic case for H_x where thus $\kappa_1(x)<0$ holds, it seems like we have to stick to the more involved formulation deduced beforehand.

On this note, we can here conclude that the above result together with the assumption (A3) and (A4) on our objective functional also imply that the regularization parameter ω remains bounded over the course of the minimization process. Let us now deduce the ensuing global convergence results for the inexact Proximal Newton method as presented in the scheme of Algorithm 11.

Algorithm 11: Inexact Proximal Newton Method

Data: Starting point $x_0 \in \operatorname{dom} g$, sufficient decrease parameter $\gamma \in]0,1]$, initial values $\omega_0 \geq 0$ and $0 \leq \eta_0 < 1$, threshold $\varepsilon > 0$ for stopping criterion
Initialization: $k = 0$;
while $\frac{1+\omega_k}{1-\eta_k} \left\| \Delta s_k(\omega_k) \right\|_X \geq \varepsilon$ **do**
  Choose $\mu \in \partial_F g(x_k)$ and compute norm term for $\tilde{\omega}$ as in (4.1.27) via the linearized minimization problem (4.1.22);
  Compute a trial step $\Delta s_k(\omega_k)$ according to (3.2.13) which suffices the inexactness criteria (4.1.14) and (4.1.27);
  while *Sufficient decrease criterion* (4.1.28) *is not satisfied* **do**
    Increase ω_k appropriately;
    Recompute trial step $\Delta s_k(\omega_k)$ as above;
  end
  Update the current iterate to $x_{k+1} \leftarrow x_k + \Delta s_k(\omega_k)$;
  Decrease ω_k appropriately to some $\omega_{k+1} < \omega_k$ for next iteration;
  Adapt η_k appropriately to some η_{k+1} for next iteration;
  Update the sequence index $k \leftarrow k + 1$;
end

For this reason, we will first prove that the right-hand side of (4.1.29), i.e., the norm of the exactly computed so-called *comparative steps* $\Delta x(\tilde{\omega}_{\max} + \omega + 1)$, converges to zero along the sequence of iterates generated by inexact updates. Here, it will come in handy to define $\omega^c \coloneqq \tilde{\omega}_{\max} + \omega + 1$ for the regularization parameter of the comparative exact update steps. Note that this quantity is in particular also bounded both from above and below.

Lemma 4.1.6: Convergence of Update Step Norms in the Inexact Case

Let $(x_k) \subset X$ be the sequence generated by the inexact Proximal Newton method globalized via (3.2.13) starting at any $x_0 \in \operatorname{dom} g$. Additionally, suppose that the subgradient inexactness criterion (4.1.25) and the sufficient decrease criterion (4.1.28) are satisfied for all $k \in \mathbb{N}$. Then, either $F(x_k) \to -\infty$ or $\left\| \Delta x_k(\omega_k^c) \right\|_X \to 0$ for $k \to \infty$.

Proof. By (4.1.29) the sequence $F(x_k)$ is monotonically decreasing. Thus, we have either $F(x_k) \to -\infty$ or $F(x_k) \to \underline{F}$ for some $\underline{F} \in \mathbb{R}$ and thereby in particular $F(x_{k+1}) - F(x_k) \to 0$. As a consequence of (4.1.29), then also $\left\| \Delta x_k(\omega_k^c) \right\|_X \to 0$ holds. □

Remark. *In particular, the above result is enabled by the uniformity of the prefactor on the right-hand side of* (4.1.29). *This estimate, on the other hand, directly follows from the sufficient decrease criterion* (4.1.28) *and the subgradient inexactness criterion* (4.1.25). *Reminiscing the corresponding argument in the exact case, we remember that there the explicit bound* (3.2.21) *was necessary in order to ensure uniformity in that regard.*

As we will see, the above formulation will also turn out to be sufficient for the proof of satisfying global convergence results. Furthermore, exactly computed update steps apparently satisfy the subgradient inexactness criterion which implies that they are in particular admissible for the inexact global convergence theory. As a consequence, the corresponding results can also be achieved for exact update steps not explicitely sufficing (3.2.21) *which played a central role in the previously considered scenario.*

As before, the above result does not comprise the convergence of the sequence of iterates itself which is desirable in the context. In the exact case of update step computation it was possible to take advantage of first order optimality conditions of the exactly solved subproblem for the actual update steps and from there achieve a proper global convergence result at least in the strongly convex case, cf. Section 3.2.3. Due to the presence of inexactness in the update step computation, this strategy has to be slightly adjusted in the current scenario, i.e., applied to the comparative update steps $\Delta x(\omega^c)$ instead. To this end, for some $k \in \mathbb{N}$ and iterate $x_k \in X$ we introduce the so-called *corresponding comparative iterate*

$$y_k \coloneqq x_k + \Delta x_k(\omega_k^c) = \mathcal{P}_g^{H_{x_k}+\omega_k^c\mathfrak{R}}\big((H_{x_k} + \omega_k^c\mathfrak{R})x_k - f'(x_k)\big)\,. \tag{4.1.33}$$

Note here that the comparative iterate uses a theoretical exact update but origins at the iterate x_k which belongs to our inexact method. Also, for every $k \in \mathbb{N}$ the identity $y_k - x_k = \Delta x_k(\omega_k^c)$ holds by the definition of y_k.

With this definition at hand, we are in the position to discuss at least subsequential convergence of our algorithm to a stationary point. In the following, we will assume throughout that the sequence of objective values $\big(F(x_k)\big)_{k\in\mathbb{N}}$ is bounded from below. Again, we start with the case of convergence in norm and the arguments here can be compared to the ones conducted for the proof of Theorem 3.2.15 but have to be applied to the comparative sequence:

Theorem 4.1.7: Stationarity of Limit Points in the Inexact Case

Assume that the subgradient inexactness criterion (4.1.25) and the sufficient decrease criterion (4.1.28) are fulfilled. Then, all accumulation points $\bar{x}$ (in norm) of the sequence of iterates (x_k) generated by Algorithm 11 are stationary points of problem (3.2.1).

Let now $(x_{k_l}) \subset (x_k)$ be the subsequence converging to $\bar{x}$. In particular, the corresponding comparative subsequence (y_{k_l}) defined via (4.1.33) satisfies

$$\operatorname{dist}\big(\partial_F F(y_{k_l}), 0\big) \to 0 \quad \text{and} \quad \big\|x_{k_l} - y_{k_l}\big\|_X \to 0\,,$$

i.e., also $y_{k_l} \to \bar{x}$ for $l \to \infty$.

Proof. We simplify notation by referring to subsequence indices k_l as k. As mentioned beforehand, for the corresponding comparative sequence (y_k) we have $y_k - x_k = \Delta x_k(\omega^c)$. Consequently, also $y_k \to \bar{x}$ holds by $\big\|\Delta x_k(\omega^c)\big\|_X \to 0$ due to Lemma 4.1.6. The proximal representation of y_k in (4.1.33) is equivalent to the minimization problem

$$y_k = \operatorname*{arg\,min}_{y\in X} g(y) + \frac{1}{2}\big(H_{x_k} + \omega_k^c\mathfrak{R}\big)(y)^2 - \big((H_{x_k} + \omega_k^c\mathfrak{R})x_k - f'(x_k)\big)y$$

which yields the first order optimality conditions given by the dual space inclusion

$$0 \in \partial_F g(y_k) + f'(x_k) + \big(H_{x_k} + \omega_k^c\mathfrak{R}\big)(y_k - x_k)\,.$$

This, on the other hand, is equivalent to

$$\big(H_{x_k} + \omega_k^c \mathfrak{R}\big)(x_k - y_k) + f'(y_k) - f'(x_k) \in \partial_F g(y_k) + f'(y_k) = \partial_F F(y_k) \tag{4.1.34}$$

the remainder term on the left-hand side of which we can estimate in norm via

$$\begin{aligned}\big\|\big(H_{x_k} + \omega_k^c \mathfrak{R}\big)(x_k - y_k) + f'(y_k) - f'(x_k)\big\|_{X^*} &\leq \big(M + \omega_k^c + L_f\big)\big\|x_k - y_k\big\|_X \\ &= \big(M + \omega_k^c + L_f\big)\big\|\Delta x_k(\omega_k^c)\big\|_X \to 0\end{aligned}$$

for $k \to \infty$ where M denotes the uniform bound on the second order bilinear form norms from assumption (A2).

In order to now achieve the optimality assertion of the accumulation point $\bar{x}$, we have to slightly adjust (4.1.34) for the use of the convex subdifferential and its direct characterization. To this end, we consider a bilinear form $Q : X \times X \to \mathbb{R}$ such that the function $\tilde{g} : X \to \mathbb{R}$ defined via $\tilde{g}(x) \coloneqq g(x) + \frac{1}{2}Q(x)^2$, $x \in X$, is convex. As usual, $Q \coloneqq H_{x_k} + \omega_k \mathfrak{R}$ accounts for a reasonable choice. Inserting such a $Q(y_k)$-term into (4.1.34) thus yields

$$\big(H_{x_k} + \omega_k^c \mathfrak{R}\big)(x_k - y_k) + f'(y_k) - f'(x_k) \in \partial \tilde{g}(y_k) + \{f'(y_k) - Q(y_k)\}$$

for the convex subdifferential of $\tilde{g}$. The left-hand side now as before converges to zero in X^* and, consequently, we know that for every $k \in \mathbb{N}$ there exists some $\tilde{\rho}_k \in \partial \tilde{g}(y_k)$ such that we can define $\tilde{\rho} \coloneqq \lim_{k\to\infty} \tilde{\rho}_k = -f'(\bar{x}) + Q\bar{x}$ by the convergence of also y_k to $\bar{x}$. The lower semi-continuity of $\tilde{g}$ together with the definition of the convex subdifferential $\partial \tilde{g}$ directly yields

$$\begin{aligned}\tilde{g}(u) - \tilde{g}(\bar{x}) &= \tilde{g}(u) - g(\bar{x}) - \frac{1}{2}Q(\bar{x})^2 \geq \tilde{g}(u) - \liminf_{k\to\infty} g(y_k) - \lim_{k\to\infty} \frac{1}{2}Q(y_k)^2 \\ &= \liminf_{k\to\infty} \tilde{g}(u) - \tilde{g}(y_k) \geq \liminf_{k\to\infty} \tilde{\rho}_k(u - y_k) = \lim_{k\to\infty} \tilde{\rho}_k(u - y_k) = \tilde{\rho}(u - \bar{x})\end{aligned}$$

for any $u \in X$ which proves the inclusion $\tilde{\rho} \in \partial \tilde{g}(\bar{x})$. The evaluation of the latter limit expression can easily be retraced by splitting

$$\tilde{\rho}_k(u - y_k) = \tilde{\rho}_k(u - \bar{x}) + (\tilde{\rho}_k - \tilde{\rho})(\bar{x} - y_k) + \tilde{\rho}(\bar{x} - y_k)\,. \tag{4.1.35}$$

In particular, we recognize $\tilde{\rho} \in \partial \tilde{g}(\bar{x})$ as $-f'(\bar{x}) + Q\bar{x} \in \partial \tilde{g}(\bar{x})$ and equivalently $-f'(\bar{x}) \in \partial_F g(\bar{x})$ for the Frechét-subdifferential ∂_F. This implies $0 \in \partial_F F(\bar{x})$, i.e., the stationarity of our accumulation point $\bar{x}$ as in Definition 3.2.5. □

In Section 3.2.3, the criterion $(1 + \omega_k)\big\|\Delta x_k(\omega_k)\big\|_X \leq \varepsilon$ for some threshold value $\varepsilon > 0$ has been used as a condition for the optimality of the current iterate up to some prescribed accuracy. Now, however, we also have to address the influence of inexactness in order to obtain similar significance for the norm of currently considered update steps $\Delta s_k(\omega_k)$. To this end, we consider

$$\begin{aligned}(1 - \eta)\big\|\Delta x(\omega)\big\|_X &\leq \big\|\Delta x(\omega)\big\|_X - \big\|\Delta x(\omega) - \Delta s(\omega)\big\|_X \\ &\leq \big\|\Delta x(\omega) - \big(\Delta x(\omega) - \Delta s(\omega)\big)\big\|_X = \big\|\Delta s(\omega)\big\|_X\end{aligned} \tag{4.1.36}$$

by (4.1.13) for $\eta < 1$. This estimate suggests that we can also here consider the adequately scaled version

$$\frac{1+\omega_k}{1-\eta_k}\left\|\Delta s_k(\omega_k)\right\|_X < \varepsilon \tag{4.1.37}$$

as the stopping criterion in the formulation and later implementations of Algorithm 11.

As has been the case in the exact scenario, the assertions from Theorem 4.1.7 can be improved further under additional structural assumptions concerning compactness and convexity. On this count, we also here note that in (4.1.35) even weak convergence of $x_k \rightharpoonup \bar{x}$ would be sufficient for the evaluation of the corresponding limit. Unfortunately, in the latter case we cannot evaluate $f'(y_k) \to f'(\bar{x})$. In order to extend our proof to this situation, we require some more structure for both of the parts of our composite objective functional. The proof is completely analogous to the one of Theorem 3.2.17.

Theorem 4.1.8: Global Convergence Under Additional Structural Assumptions in the Inexact Case

Let f be of the form $f(x) = \hat{f}(x) + \check{f}(Kx)$ where K is a compact operator. Additionally, assume that $g + \hat{f}$ is convex and weakly lower semi-continuous in a neighborhood of stationary points of (3.2.1). Suppose that $\check{f}$ satisfies the assumptions made on f beforehand. Then, weak convergence of the inexact sequence of iterates $x_k \rightharpoonup \bar{x}$ suffices for $\bar{x}$ to be a stationary point of (3.2.1).

If F is strictly convex and radially unbounded, the whole sequence (x_k) converges weakly to the unique minimizer x_* of F. If F is κ-strongly convex, with $\kappa > 0$, then $x_k \to x_*$ in norm.

Even though it might look irritating at first glance, after taking a second thought it is not too surprising that the global convergence results we have achieved here all in all mirror the ones from the exact case in Section 3.2.3: With the aid of the comparative sequence introduced in (4.1.33), we have established the estimate that globally our inexact updates achieve the same amount of progress in view of the overall minimization as an exact variant of the Proximal Newton method investigated before, just using a (probably unreasonably) high regularization parameter ω_k^c.

The global convergence for this comparative sequence of iterates now behaves exactly as we have observed before. Thus, since global convergence theory in general does not use quantifiable measures which incorporate whether the regularization parameter is chosen 'tightly' or not but only states whether the corresponding sequence converges or not, exactly these global convergence results carry over to the inexact case here. All we had to ensure is the existence of the estimate with respect to such a highly regularized comparative sequence. This is what has motivated the subgradient inexactness criterion and it has done the job just fine.

4.1.4 Transition to Fast Local Convergence

Another similarity to the convergence analysis in the exact case is that also here, we have to manage the transition from the globalization phase to the local convergence phase in order to then benefit from the local acceleration result in Theorem 4.1.4. To this end, we have to again make sure that – at least close to stationary points of (3.2.1) – arbitrarily small regularization parameters $\omega \geq 0$ yield update steps that give us sufficient decrease in F according to the criterion formulated in (4.1.28).

As a starting point, a rather technical auxiliary result is required. Similar to the formulation of Corollary 3.2.27, it sets the limit behavior of inexact update steps in relation with the distance of consecutive iterates to the minimizer of (3.2.1):

Corollary 4.1.9: Limit Behavior of Damped Inexact Update Steps

Let x and $x_+(\omega) = x + \Delta s(\omega)$ be two consecutive iterates with update step $\Delta s(\omega)$ sufficing (4.1.13) for some $0 \leq \eta < 1$. Furthermore, consider an optimal solution x_* of (3.2.1) sufficiently close to which x and $x_+(\omega)$ are located.

Then, the following estimates eventually hold for $\kappa_1(x) + \kappa_2 > 0$:

$$\big\|x_+(\omega) - x_*\big\|_X \leq (3 + 2\eta)\big\|x - x_*\big\|_X \quad \text{and} \quad \big\|x - x_*\big\|_X \leq \frac{2}{1-\eta}\big(1 + \frac{\omega}{\kappa_1 + \kappa_2}\big)\big\|\Delta s(\omega)\big\|_X \,.$$

Remark. *In particular, these eventual norm estimates have implications on the limit behavior of the respective terms. If we now have $\xi = o\big(\|x_+(\omega) - x_*\|_X\big)$ for some $\xi \in X$, then $\xi = o\big(\|x - x_*\|_X\big)$ immediately holds and from there we obtain $\xi = o\big(\|\Delta s(\omega)\|_X\big)$ in the same way.*

Proof. Our proof here mainly exploits the local superlinear convergence of exactly computed and undamped update steps $\Delta x := \Delta x(0)$ from Theorem 4.1.4 and then uses the respective estimates in order to introduce the influences of both damping and inexactness. For the first asserted estimate, we take a look at

$$\begin{aligned}\big\|x_+(\omega) - x_*\big\|_X &\leq \big\|x - x_*\big\|_X + \big\|\Delta s(\omega)\big\|_X \leq \big\|x - x_*\big\|_X + (1+\eta)\big\|\Delta x\big\|_X \\ &\leq (2+\eta)\big\|x - x_*\big\|_X + (1+\eta)\big\|x + \Delta x - x_*\big\|_X\end{aligned}$$

where the second step involved (4.1.30) together with $\big\|\Delta x(\omega)\big\|_X \leq \big\|\Delta x\big\|_X$ as proven in Lemma 3.2.26. From here, we use the superlinear convergence of exact updates in the form of the existence of some modulus of continuity $\psi : [0, \infty[\to [0, \infty[$ with $\psi(t) \to 0$ for $t \to 0$ such that

$$\big\|x + \Delta x - x_*\big\|_X = \psi\big(\big\|x - x_*\big\|_X\big)\big\|x - x_*\big\|_X$$

holds in the limit of $x \to x_*$. Thus, we obtain

$$\big\|x_+(\omega) - x_*\big\|_X \leq \big[2 + \eta + (1+\eta)\psi\big(\big\|x - x_*\big\|_X\big)\big]\big\|x - x_*\big\|_X \leq (3 + 2\eta)\big\|x - x_*\big\|_X$$

since eventually we can assume the ψ-term to be smaller than one in that same limit. This completes the proof of the first asserted estimate.

For the second one, we take advantage of

$$\big\|\Delta x\big\|_X \leq \big(1 + \frac{\omega}{\kappa_1(x) + \kappa_2}\big)\big\|\Delta x(\omega)\big\|_X$$

from Lemma 3.2.26 together with again the superlinear convergence as above and find that

$$\begin{aligned}\big\|x - x_*\big\|_X &\leq \big\|x + \Delta x - x_*\big\|_X + \big\|\Delta x\big\|_X \\ &\leq \psi\big(\big\|x - x_*\big\|_X\big)\big\|x - x_*\big\|_X + \big(1 + \frac{\omega}{\kappa_1(x) + \kappa_2}\big)\big\|\Delta x(\omega)\big\|_X\end{aligned}$$

holds. Since the ψ-term eventually will be smaller than one half, from here we infer

$$\big\|x - x_*\big\|_X \leq \frac{1 + \frac{\omega}{\kappa_1(x)+\kappa_2}}{1 - \psi\big(\big\|x - x_*\big\|_X\big)} \big\|\Delta x(\omega)\big\|_X \leq 2\big(1 + \frac{\omega}{\kappa_1(x) + \kappa_2}\big) \big\|\Delta x(\omega)\big\|_X \,.$$

The inexactness of update step computation now enters the above estimate using the inequality $\big\|\Delta x(\omega)\big\|_X \leq \frac{1}{1-\eta} \big\|\Delta s(\omega)\big\|_X$ from (4.1.36). This completes the proof of the corollary. □

In what follows, it will again be important several times that the second order bilinear forms H_x satisfy a bound of the form

$$\big(H_{x_+(\omega)} - H_x\big)\big(x_+(\omega) - x_*\big)^2 = o\big(\big\|x - x_*\big\|_X^2\big) \quad \text{for } x \to x_* \,. \tag{4.1.38}$$

As we have pointed out in the formulation of Proposition 3.2.28, it is easy to see that the bound holds if we either have uniform boundedness of the second order bilinear forms together with superlinear convergence of the iterates or continuity of the mapping $x \mapsto H_x$ together with mere convergence of the iterates to x_*. In our scenario, we conclude that according to Theorem 4.1.4 it is sufficient that both the regularization parameters $\omega_k \geq 0$ and the forcing terms $\eta_k \geq 0$ converge to zero as we approach a stationary point $x_* \in X$ of (3.2.1) together with assumption (A2) from the introductory section. We will later on establish this convergence of (ω_k) and (η_k) in the specific implementation of our algorithm.

With the auxiliary estimates from Corollary 4.1.9 and Lemma 3.2.26 together with the thoroughly discussed additional assumption from (4.1.38) at hand, we can now turn our attention to the actual admissibility of arbitrarily small regularization parameters close to optimal solutions of (3.2.1).

For that matter, also here we furthermore suppose f to be second order semi-smooth at stationary points x_* of (3.2.1) with respect to the mapping $H\colon X \to \mathcal{L}^{(2)}(X, \mathbb{R}), x \mapsto H_x$, which we remember to be represented by the estimate

$$f(x_* + \xi) = f(x_*) + f'(x_*)\xi + \frac{1}{2} H_{x_*+\xi}(\xi, \xi) + o\big(\big\|\xi\big\|_X^2\big) \quad \text{for } \big\|\xi\big\|_X \to 0 \,. \tag{4.1.39}$$

With these two crucial assumptions back on our mind, we can now take a look at the generalization of the admissibility of damped update steps to the inexact case. Again, the result takes a similar look as before but the proof has to be adapted to the less generously formulated scenario here.

Proposition 4.1.10: Admissibility of Inexact Update Steps Close to Stationary Points

Suppose that the additional assumptions (4.1.38) and (4.1.39) hold. Furthermore, assume that the update steps $\Delta s(\omega)$ computed as inexact solutions of (3.2.13) at iterates $x \in X$ for some $\omega \geq 0$ satisfy the inexactness criteria (4.1.13) for $0 \leq \eta < 1$ and (4.1.25) for $\tilde{\omega}_{\max} > 0$. Let $x_* \in X$ be an optimal solution of (3.2.1) near which $\kappa_1(x) + \kappa_2 > 0$ holds.

Then, for any $\gamma \in]0, 1]$, we can find a neighborhood $U_{\omega,\gamma} \subset X$ of x_* such that at all $x \in U_{\omega,\gamma}$ the update $\Delta s(\omega)$ is admissible for sufficient decrease according to (4.1.28) for that γ.

Proof. We take a look back at the proof of Proposition 3.2.28 and employ the same telescoping strategy in order to obtain

$$\begin{aligned} f\big(x_+(\omega)\big) - f(x) - f'(x)\Delta s(\omega) - \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2 \\ = \left[f\big(x_+(\omega)\big) - f(x_*) - f'(x_*)\big(x_+(\omega) - x_*\big) - \frac{1}{2}H_{x_+(\omega)}\big(x_+(\omega) - x_*\big)^2\right] \\ - \left[f(x) - f(x_*) - f'(x_*)(x - x_*) - \frac{1}{2}H_x(x - x_*)^2\right] \\ - \big(f'(x) - f'(x_*) - H_x(x - x_*)\big)\Delta s(\omega) + \frac{1}{2}(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2 \end{aligned}$$

where we can again use the second order semi-smoothness of f according to (4.1.39) for the first two terms as well as the semi-smoothness of f' as in (3.2.11) for the third one. This implies

$$\begin{aligned} f\big(x_+(\omega)\big) - f(x) - f'(x)\Delta s(\omega) - \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2 = o\big(\|x_+(\omega) - x_*\|_X^2\big) + o\big(\|x - x_*\|_X^2\big) \\ + o\big(\|x - x_*\|_X\big)\|\Delta s(\omega)\|_X + \rho(x,\omega) \end{aligned}$$

where we denoted $\rho(x,\omega) \coloneqq \frac{1}{2}(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2$. Due to the limit behavior of inexact update step norms investigated over the course of Corollary 4.1.9, this yields

$$f\big(x + \Delta s(\omega)\big) - f(x) - f'(x)\Delta s(\omega) - \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2 = \rho(x,\omega) + o\big(\|\Delta s(\omega)\|_X^2\big)\,. \tag{4.1.40}$$

With these preliminary estimates deduced, we can now choose some $\gamma \in]0,1]$ and $\omega \geq 0$ for the combination of which the sufficient decrease criterion (4.1.28) has to be satisfied. To this end, we also here define the decrease ratio function

$$\gamma\colon X \times [0,\infty[\to]-\infty,\infty]\,, \quad \gamma(x,\omega) \coloneqq \frac{F\big(x + \Delta s(\omega)\big) - F(x)}{\lambda_{x,\omega}\big(\Delta s(\omega)\big)}$$

which should be larger than the $\gamma \in]0,1]$ chosen beforehand for $\Delta s(\omega)$ to yield sufficient decrease. In order to prove the existence of a neighborhood $U_{\gamma,\omega}$ in which inexact update steps do so, we have to investigate the limit behavior of $\gamma(x,\omega)$ for the chosen regularization parameter ω in the limit of $x \to x_*$. The identity (4.1.40) from above now provides us with

$$F\big(x + \Delta s(\omega)\big) - F(x) = \lambda_{x,\omega}\big(\Delta s(\omega)\big) - \frac{\omega}{2}\|\Delta s(\omega)\|_X^2 + \rho(x,\omega) + o\big(\|\Delta s(\omega)\|_X^2\big)$$

which we insert into the decrease ratio function from above and estimate

$$\begin{aligned} \gamma(x,\omega) &= 1 + \frac{-\frac{\omega}{2}\|\Delta s(\omega)\|_X^2 + \rho(x,\omega) + o\big(\|\Delta s(\omega)\|_X^2\big)}{\lambda_{x,\omega}\big(\Delta s(\omega)\big)} \\ &= 1 + \frac{\frac{\omega}{2}\|\Delta s(\omega)\|_X^2 - \rho(x,\omega) - o\big(\|\Delta s(\omega)\|_X^2\big)}{\big|\lambda_{x,\omega}\big(\Delta s(\omega)\big)\big|} \end{aligned} \tag{4.1.41}$$

since from the computation strategy for $\Delta s(\omega)$ we in particular have $\lambda_{x,\omega}\big(\Delta s(\omega)\big) \leq 0$ by (4.1.31). As we take a closer look at the latter estimate, we in particular recognize that the absolute value satisfies

$$\big|\lambda_{x,\omega}\big(\Delta s(\omega)\big)\big| \geq \frac{\big(\omega + \kappa_1(x) + \kappa_2\big)^2}{2(1+\eta)^2\big(\tilde{\omega}_{\max} + \omega + \kappa_1(x) + \kappa_2\big)}\|\Delta s(\omega)\|_X^2 =: \frac{C}{2}\|\Delta s(\omega)\|_X^2 \tag{4.1.42}$$

where $C = C(\tilde{\omega}_{\max}, \omega + \kappa_1(x) + \kappa_2, \eta) > 0$ remains bounded in the limit of $\omega \to 0$ and is also well-defined in the limit case of $\omega = 0$ close to stationary points x_* with $\kappa_1(x) + \kappa_2 > 0$ for x near x_*.

We may assume that the numerator of the latter expression in (4.1.41) is non-positive, otherwise the desired inequality for $\gamma(x, \omega)$ is trivially fulfilled. Thus, we take advantage of (4.1.42) in order to decrease the positive fraction term and thus achieve

$$\gamma(x, \omega) \geq 1 + \frac{\omega}{C} - \frac{o\big(\big\|\Delta s(\omega)\big\|_X^2\big) + \rho(x, \omega)}{\frac{C}{2}\big\|\Delta s(\omega)\big\|_X^2}\,.$$

Now, the assumption (4.1.38) for the ρ-term together with the limit behavior estimates from Corollary 4.1.9 immediately implies that we can find a sufficiently small neighborhood of x_* such that

$$\frac{o\big(\big\|\Delta s(\omega)\big\|_X^2\big) + \rho(x, \omega)}{\big\|\Delta s(\omega)\big\|_X^2} < \frac{1}{2}[C(1 - \gamma) + \omega]$$

holds in that neighborhood as which we in particular choose the desired $U_{\gamma,\omega}$ from the assertion. By construction $\Delta s(\omega)$ is thus admissible for sufficient decrease according to (4.1.28) for that γ if it has been computed at $x \in U_{\gamma,\omega}$. □

On a last remark, we note that the convergence of the decrease ratio function to something greater equal than one as formulated in (3.2.39) pertains to hold also in the inexact scenario. As we have mentioned before, this will contribute to the motivation and investigation of parameter choice strategies later on in Section 4.2.

4.1.5 The Alternative Sufficient Decrease Criterion in the Inexact Case

Another algorithmic development which has to be adapted to the inexact computation of update steps is the alternative sufficient decrease criterion from Section 3.2.6. While it has originally been formulated using a Fréchet subderivative $\mu_+ \in \partial_F g\big(x + \Delta x(\omega)\big)$ at the updated iterate, the optimality of update steps $\Delta x(\omega)$ from (3.2.13) provided us with the computationally and analytically advantageous reformulation (3.2.42). In our present scenario, where update steps are not computed exactly in order to spare computational effort, the same characterization of such subderivatives is not accessible.

Fortunately, we have already clarified the roles of the respective formulations (3.2.40) and (3.2.42) of the alternative sufficient decrease criterion within the proofs of the corresponding results in Section 3.2.6. We remember that the direct formulation via the subderivative $\mu_+ \in \partial_F g\big(x + \Delta x(\omega)\big)$ was rather of motivational character and, thus, for the decrease parameter $\gamma \in]0, 1]$ from before, we continue to use

$$\big[f'\big(x + \Delta s(\omega)\big) - f'(x) - H_x\big(\Delta s(\omega)\big)\big]\Delta s(\omega) \leq \frac{1 - \gamma}{2}\omega\big\|\Delta s(\omega)\big\|_X^2 \qquad (4.1.43)$$

from Definition 3.2.29 as the alternative sufficient decrease criterion for an inexactly computed update step $\Delta s(\omega)$ close to optimal solutions of (3.2.1). In particular, directly using the "reformulated" version as a definition here frees us from the dilemma of having to choose some subderivative at the inexactly updated iterate which is not given by a fixed expression stemming from optimality of the update step. This strategy yields an algorithmically efficient

and in sufficient generality implementable routine for which we do not have to manually pick a subderivative and ensure crucial estimates for it.

Even though apparently the ultimate goal here is to ensure global "residual" convergence from our alternative sufficient decrease criterion (4.1.43), we will proceed as we did in Section 3.2.6 and first ensure satisfiability for sufficiently large ω:

Lemma 4.1.11: Satisfiability of the Alternative Sufficient Decrease Criterion in the Inexact Case

The alternative sufficient decrease criterion (4.1.43) is satisfied for $\gamma \in]0, 1]$ if the regularization parameter ω suffices the estimate

$$\omega \geq \frac{2(L_f + M)}{1 - \gamma}\,.$$

Proof. The proof here follows along the same lines as the one of Lemma 3.2.30 in the exact case. In particular, there we did not at all use the optimality of update steps $\Delta x(\omega)$ but only the Lipschitz continuity of f' together with the boundedness of the second order bilinear forms H_x. □

In order to now conclude convergence of update step norms to zero and thereby, as usually, global convergence results of our inexact Proximal Newton method, we have to reconsider the global convergence strategy in the inexact case in particular. Before, in Section 4.1.3, we have introduced the subgradient inexactness criterion (4.1.25) in order to deduce global convergence results from the original sufficient decrease criterion (4.1.28). For this reason, this second inexactness criterion becomes somewhat insignificant for our intention of developing an alternative global convergent theory close to optimal solutions. This conclusion leaves us with the freedom to choose another condition which qualifies update steps to be computed accurately enough for our minimization strategy.

It will turn out that it is sufficient to demand a sligthly different decrease estimate within the modified second order model $\lambda_{x,\omega} \colon X \to]-\infty, \infty]$ from (3.2.12) in the form of

$$\lambda_{x,\omega}\big(\Delta s(\omega)\big) \leq \eta_{\mathrm{glob}} \big\|\Delta s(\omega)\big\|_X^2 \tag{4.1.44}$$

for the *global forcing term* $\eta_{\mathrm{glob}} \in]-\frac{1}{2}\big(\omega + \kappa_1(x) + \kappa_2\big), \frac{\gamma}{2}\omega[$ and our inexact candidate $\Delta s(\omega)$ computed according to (3.2.13). We will explain the choice of the range of η_{glob} above after its use within the proof of global residual convergence in the next proposition. Before verifying this result, we remark that, due to the theoretical bound (3.2.16), satisfiability of (4.1.44) is given as soon as we can assume convergence of objective values within the subproblem solver. As noted beforehand for the satisfiability of the subgradient inexactness criterion, this might depend on continuity assumptions of g but we will not elaborate on this demand here.

The following global convergence result again features the assumption of the current update step already being sufficiently small close to the optimal solution of the underlying minimization problem. This rather obscure formulation will at least be somewhat clarified over the course of the proof of the statement.

Proposition 4.1.12: Global Residual Convergence using the Alternative Decrease Criterion in the Inexact Case

Let f' be semi-smooth at a stationary point $x_* \in X$ with respect to the mapping $H\colon X \to \mathcal{L}^{(2)}(X,\mathbb{R})$ which is continuous at x_*. Suppose that inexact update steps $\Delta s(\omega)$ are computed according to (3.2.13) such that they suffice (4.1.13) for some $0 \leq \eta < 1$ and (4.1.44) for $\eta_{\text{glob}} < \frac{\gamma}{2}\omega$.

Then, if update steps are already sufficiently small when the alternative sufficient decrease criterion (3.2.40) is used, global convergence from the original algorithm is continued insofar that we now have $\Delta x_k(\omega) \to 0$ for $k \to \infty$ (in case F is bounded from below) together with all ensuing global convergence results from Section 4.1.3.

Proof. We use the sufficient second inexactness criterion (4.1.44) insofar that

$$g\big(x+\Delta s(\omega)\big) - g(x) \leq -f'(x)\Delta s(\omega) - \frac{1}{2}\big(H_x + \omega\mathfrak{R}\big)\big(\Delta s(\omega)\big)^2 + \eta_{\text{glob}}\omega\big\|\Delta s(\omega)\big\|_X^2$$

holds which we – as in the proof of Proposition 3.2.31 – turn into

$$\begin{aligned} &F\big(x+\Delta s(\omega)\big) - F(x) \\ &\quad = \big[f\big(x+\Delta s(\omega)\big) - f(x) + \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2\big] - \big[f'(x) + H_x\big(\Delta s(\omega)\big)\big]\Delta s(\omega) \\ &\qquad - \frac{\omega}{2}\big\|\Delta s(\omega)\big\|_X^2 + \eta_{\text{glob}}\big\|\Delta s(\omega)\big\|_X^2 . \end{aligned} \tag{4.1.45}$$

For the first bracket term, we apply the same telescoping strategy as before and by Corollary 4.1.9 thus obtain

$$f\big(x+\Delta s(\omega)\big) - f(x) + \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2 = f'\big(x+\Delta s(\omega)\big)\Delta x(\omega) + o\big(\big\|\Delta s(\omega)\big\|_X^2\big)$$

for $x \to x_*$. As we insert this finding into (4.1.45), our alternative sufficient decrease criterion (4.1.43) in this scenario then – similarly as before – in the same limit provides us with

$$F\big(x+\Delta s(\omega)\big) - F(x) \leq -\big(\frac{\gamma}{2}\omega - \eta_{\text{glob}}\big)\big\|\Delta s(\omega)\big\|_X^2 + o\big(\big\|\Delta s(\omega)\big\|_X^2\big) \leq -c\big\|\Delta s(\omega)\big\|_X^2$$

by our choice of $\eta_{\text{glob}} < \frac{\gamma}{2}\omega$ and the fact that our current iterate x is already located sufficiently close to an the optimal solution x_*. For F bounded from below, this lets us conclude the convergence of inexact update step norms to zero which can be transferred to their exact counterparts via (4.1.36) with $\eta < 1$.

Thus, we have the convergence to zero of exact update step norms along our inexactly computed sequence of iterates. This circumstance provides us with a comparative sequence as in Section 4.1.3 and thus allows us to deduce the same global convergence results also here. □

Remark. *The comparative sequence here employs the original regularization parameter ω in contrast to the former case in Section 4.1.3 where the adapted version ω^c has been used.*

The proof of the above global convergence result now lets us more elaborately explain the range of the global forcing term $\eta_{\text{glob}} \in \,]-\frac{1}{2}\big(\omega+\kappa_1(x)+\kappa_2\big), \frac{\gamma}{2}\omega[$: While theoretical satisfiability of the corresponding inexactness criterion (4.1.44) determines the lower end of

the interval, the upper bound on it is fixed by the arguments above. We still require a negative $\|\Delta s(\omega)\|_X^2$-term in addition to the o-terms arising from our semi-smoothness assumptions on f and for this reason need $\eta_{\text{glob}} < \frac{\gamma}{2}\omega$.

The choice of the forcing term η_{glob} within this range can also be discussed quite intuitively: While large values make the inexactness criterion (4.1.44) rather "complaisant", they also diminish the absolute value of said negative prefactor within the proof of global residual convergence. This choice also diminishes the unspecified bound on the norm of update steps which allows the theory for global convergence deduced above to start operating. In conclusion, large values of η_{glob} relax the alternative second inexactness criterion with the trade-off that we take some risk in global residual convergence. A small choice on the other hand reverts this effect: While the inexactness criterion becomes harder to satisfy and we thus have to invest additional computational effort to meet the demand, we gain safety in residual global convergence. However, it is reasonable to choose $\eta_{\text{glob}} \geq 0$ since the original lower bound $-\frac{1}{2}\big(\omega + \kappa_1(x) + \kappa_2\big)$ is generally not accessible and non-uniform within our algorithm. A non-negative value of the global forcing term particularly unlocks the implication of the alternative second inexactness criterion (4.1.44) from the original subgradient inexactness criterion (4.1.25) due to the norm estimate (4.1.31). This means that once an inexact update step has been computed accurately enough in order to suffice (4.1.25), also (4.1.44) holds for it. This will enable the final algorithmic strategy explained at the end of this section and described in the scheme of Algorithm 12.

Now that we have also in this case established a working global residual convergence theory with the aid of the alternatively defined second inexactness criterion (4.1.44), it still remains to generalize the transition result from Proposition 3.2.32 to the inexact scenario. Fortunately, also this statement can be carried over in an analogous way:

Proposition 4.1.13: Admissibility Close to Optimal Solutions Using the Alternative Sufficient Decrease Criterion in the Inexact Case

Let $f' \colon X \to X^*$ be semi-smooth at an optimal solution $x_* \in X$ of (3.2.1) with respect to the mapping $H \colon X \to \mathcal{L}^{(2)}(X, \mathbb{R})$ which is continuous at x_*.

Then, for any value $\omega > 0$, we can find a neighborhood U_ω of x_* such that the alternative sufficient decrease criterion (3.2.40) is satisfied at any $x \in U_\omega$ for that ω.

Proof. We use exactly the same telescoping strategy as in the proof of the corresponding result in the exact case from Proposition 3.2.32. The argumentation there did not involve the optimality of update steps which makes it applicable also in the inexact scenario here. We can simply reproduce this proof by referring to Corollary 4.1.9 instead of Corollary 3.2.27 for the limit behavior of inexact opposed to exact update steps. □

As before, we can also here interpret the admissibility result from above as a corresponding convergence result for an adequately defined alternative decrease ratio function. In particular, we can thus conclude that the convergence of the alternative decrease ratio function $\tilde{\gamma} \colon X \times [0, \infty[\to \mathbb{R}$ defined in (3.2.43) carries over to the inexact scenario here. Just like for the respective result in previous sections, we also remark here that we will take advantage of this perception later on when considering parameter choice strategies in Section 4.2.

Algorithmic Strategy

Also in the inexact case, we will try to take advantage of the richer global convergence theory deduced in the framework of the original sufficient decrease criterion as long as we can and only use the alternative formulation in case we suspect lacking admissibility of an update step to stem from numerical instability.

Algorithm 12: Inexact Proximal Newton Method Modified for Numerical Stability

Data: Starting point $x_0 \in X$, sufficient decrease parameter $\gamma \in]0, 1]$, initial values $\omega_0 \geq 0$ and $\eta_0 \geq 0$, thresholds $0 < \varepsilon < \tilde{\varepsilon}$ for the stopping and proximity criterion
Initialization: $k = 0$;
while $\frac{1+\omega_k}{1-\eta_k} \left\| \Delta s_k(\omega_k) \right\|_X \geq \varepsilon$ **do**
  Choose $\mu \in \partial_F g(x_k)$ and compute norm term for $\tilde{\omega}$ as in (4.1.27) via the linearized minimization problem (4.1.22);
  Compute a trial step $\Delta s_k(\omega_k)$ according to (3.2.13) which suffices the inexactness criteria (4.1.14) and (4.1.27);
  if $\frac{1+\omega_k}{1-\eta_k} \left\| \Delta x_k(\omega_k) \right\|_X \geq \tilde{\varepsilon}$ **then**
    while *Sufficient decrease criterion* (4.1.28) *is not satisfied* **do**
      Increase ω_k appropriately;
      Recompute trial step $\Delta s_k(\omega_k)$ as above;
    end
  else
    while *Sufficient decrease criteria* (4.1.28) *and* (4.1.43) *are not satisfied* **do**
      Increase ω_k appropriately;
      Recompute trial step $\Delta s_k(\omega_k)$ as above;
    end
  end
  Update the current iterate to $x_{k+1} \leftarrow x_k + \Delta s_k(\omega_k)$;
  Decrease ω_k appropriately to some $\omega_{k+1} < \omega_k$ for next iteration;
  Adapt η_k appropriately to some η_{k+1} for next iteration;
  Update the sequence index $k \leftarrow k + 1$;
end

For this reason, we do not want to decide ahead of step computation which sufficient decrease criterion we will use in order to evaluate its descent properties. We will not prescribe a fixed area where the alternative sufficient decrease criterion (4.1.43) together with the corresponding alternative second inexactness criterion (4.1.44) is used but, as discussed above, take advantage of the fact that our original subgradient inexactness criterion (4.1.25) in particular implies the alternative formulation with $\eta_{\text{glob}} = 0$ which suffices for global residual convergence. Even though this procedure might be more restrictive in update step computation as minimally necessary, it is advantageous insofar that it maintains the possibility of using the global convergence theory of the original formulation and gives additional safety when using the alternative strategy.

With these deliberations made, the algorithmic strategy of the modified inexact Proximal

Newton method combines the benefits of the exact stable version from Algorithm 10 and the inexactness in update step computation from Algorithm 11. In particular, also the proximity criterion generalizes to the inexact formulation via

$$\frac{1+\omega_k}{1-\eta_k}\big\|\Delta x_k(\omega_k)\big\|_X < \tilde{\varepsilon} \tag{4.1.46}$$

with threshold value $\tilde{\varepsilon} > 0$. For the convenience of the reader, we summarize it within the scheme of Algorithm 12. For a more illustrative overview, we again refer to the algorithmic conclusion in Section 4.3 where the final form of the algorithm is presented in Figure 4.11.

4.1.6 Numerical Results

Let us now showcase the functionality of our inexact Proximal Newton method and also compare its performance to the case of exact computation of update steps which we have investigated in Section 3.2.7. In order to make the influence of inexactness more clearly visible, we have decided to enhance the function space problem from Section 3.2.7 such that update step subproblems are harder to solve and thus it takes more TNNMG steps in order to find an exact solution.

The Objective Functional

To this end, we now consider the following function space problem on $\Omega := [0,1]^3 \subset \mathbb{R}^3$: Instead of finding a scalar function, we expanded the problem to finding a vector field

$$u \in H^1_{\Gamma_D}(\Omega,\mathbb{R}^3) := \big\{v \in H^1(\Omega,\mathbb{R}^3) \,\big|\, v = 0 \text{ on } \Gamma_D\big\}$$

where the Dirichlet boundary is given by $\Gamma_D := \{0\}\times[0,1]\times[0,1]$. The solution which we are looking for minimizes the composite objective functional F defined via

$$F(u) := f(u) + \int_\Omega c\big\|u\big\|_2 \,\mathrm{d}x \tag{4.1.47}$$

for again some parameter $c > 0$ as a weight of the Euclidean L_2-norm term where the smooth part $f\colon H^1_{\Gamma_D}(\Omega,\mathbb{R}^3) \to \mathbb{R}$ is now given by

$$f(u) := \int_\Omega \frac{1}{2}\big\|\nabla u\big\|_F^2 + \alpha\max\big(\big\|\nabla u\big\|_F - 1, 0\big)^2 + \beta\frac{u_1^3u_2^2u_3}{1+u_1^2+u_2^2+u_3^2} + \rho\cdot u\,\mathrm{d}x$$

with parameters $\alpha,\beta \in \mathbb{R}$ as well as a force field $\rho\colon \Omega \to \mathbb{R}^3$. The norm $\big\|\cdot\big\|_F$ denotes the Frobenius norm of the respective Jacobian matrices ∇u. We can see that – in contrast to (3.2.45) – we additionally replaced the simple u^3-term with the more involved fraction-term above in order to make subproblems harder to solve while preserving the differentiability properties of the smooth part f.

This slight modification does not repair the minor flaw from (3.2.45) that f technically does not satisfy the assumptions made on the smooth part of the composite objective functional specified above in the case $\alpha \neq 0$ due to the lack of semi-smoothness of the corresponding squared max-term. As we have pointed out before, we think that also here slightly going beyond the framework of theoretical results for numerical investigations can be instructive.

Similar to the respective choice in Section 3.2.7, we will choose the force-field ρ to be constant on Ω and to this end introduce the so-called load factor $\tilde{\rho} > 0$ which then determines $\rho = \tilde{\rho}(1,1,1)^T$. Again, for the sake of simplicity, we will refer to this load factor as ρ. Now that we have fully prescribed the composite objective functional F, we recognize that its non-smooth part g is again merely given by the integrated Euclidean L_2-norm term with constant prefactor $c > 0$.

Specifics of the Implementation

As far as the specifics of our implementation are concerned, they are generally similar to the ones from Section 3.2.7: We use automatic differentiation by adol-C in order to establish the second order model and TNNMG to solve update step computation subproblems. Additionally, the subproblem solver is provided with stopping criteria in the form of our inexactness criteria (4.1.14) and (4.1.27) with corresponding parameters $\eta_k \in [0, 1[$ for each iteration and global $\tilde{\omega}_{\max} > 0$.

Another topic of interest concerning the implementation of our algorithm is the choice of the aforementioned parameters ω, η and $\tilde{\omega}_{\max}$ governing the convergence behavior of our method. While – as discussed in its introduction in (4.1.27) – $\tilde{\omega}_{\max}$ can be chosen constant and is supposed to be very large, this is not the case for the regularization parameters ω and the forcing terms η. Adaptive choices for these quantities will be investigated in Section 4.2 and in the current state of our development we want to focus on the aspect of and criteria for inexactness itself. Thus, we decided to take the rather heuristic approach for ω presented in Section 3.2.7.

Similarly, we multiply the forcing term η by 0.6 for accepted updates and leave it as it is in case the increment has been rejected by the sufficient decrease criterion. This rather simple strategy for the choice of parameters ensures the convergence of both η and ω to zero along the sequence of iterates and thus also from a theoretical standpoint enables superlinear convergence as formulated in Theorem 4.1.4. For the constant determining the subgradient inexactness criterion, we decided to choose $\tilde{\omega}_{\max} = 10^{10}$.

The stopping and proximity criterion for our algorithm take into account both the regularization parameter and the forcing term as formulated in (4.1.37) and (4.1.46). The respective threshold values are identical to the ones previously used in Section 3.2.7, i.e., we choose $\varepsilon = 10^{-10}$ and $\tilde{\varepsilon} = 10^{-4}$.

Test Scenarios

Let us now consider the actual tests which we have performed in order to display the performance of our algorithm: Firstly, we will demonstrate the consistency between results of the inexact method and the exact version the functionality of which has been thoroughly investigated in Section 3.2.7. More precisely, exactly computing update steps means neglecting the additionally introduced inexactness criteria, and computing steps up to numerical accuracy in TNNMG. We remember that there, a relative norm threshold for increments is considered as a stopping criterion.

Afterwards, we exhibit the gains in effectivity by enhancing the exact algorithm with the inexactness criteria introduced above. Lastly, we analyze the implementation of the latter criteria and try to get a grasp on how they affect the process of solving the subproblem for update step computation. All results within the current section have been computed after

conducting three uniform grid refinements of the cubical domain Ω which results in $8^4 = 4096$ grid elements.

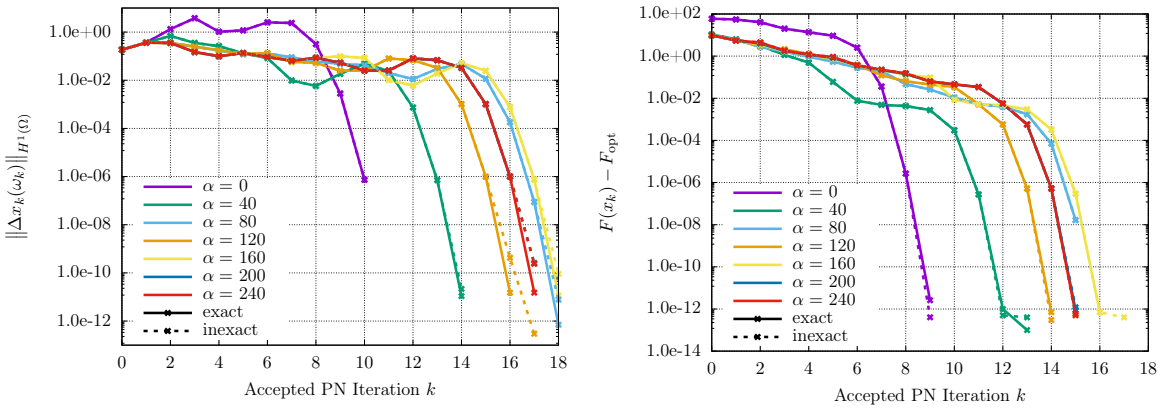
All in all, we use (4.1.47) with fixed parameters $c = 10$, $\beta = 10$, $\rho = -20$ and let $\alpha \geq 0$ vary. As we have already learned in the exact case in Section 3.2.7, increasing α magnifies the influence of the squared max-term in (4.1.47) and thus makes the corresponding minimization problem harder to solve.

Equivalence of Computed Solutions

This effect already becomes apparent in Figure 4.1a where update step norms for accepted iterates are depicted for both the exact and inexact version of our method. Here, the respective quantities for the exact procedure are represented by full lines while dashed lines signify inexact computation of update steps. Together with the plot of energy differences to the optimal value from Figure 4.1b, this in particular suggests the equivalence of results achieved by both variants of the Proximal Newton algorithm. This expectation is validated by the computation of the relative error across all grid points y^i of our discretization via the straight-forward formula

$$\mathrm{err}_{\mathrm{rel}}(y^i) \coloneqq \frac{\left\| u_{\mathrm{ex}}(y^i) - u_{\mathrm{inex}}(y^i) \right\|_2}{\left\| u_{\mathrm{ex}}(y^i) \right\|_2}$$

where we denoted by u_{ex} or u_{inex} the respective results of the exact or inexact method. In our simulations, this relative error expression reveals that the maximal discrepancy between the solutions found by the respective methods is even below numerical accuracy and yields zero in computational evaluation.



(a) Correction Norms $\left\|\Delta x_k(\omega_k)\right\|_{H^1(\Omega)}$.

(b) Energy Differences $F(x_k) - F_{\mathrm{opt}}$.

Figure 4.1: Graphs of correction norms and energy differences to the optimal value for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$ for the exact and inexact Proximal Newton method. Correction norm plots are not extended as in Section 3.2.7 for the sake of perspicuity.

Improvements in Computational Efficiency

With the validation that the inexact variant of our Proximal Newton method achieves the same solution and general convergence behavior as the exact method at hand, we can now

turn our attention to the actual reason for which we have made the deliberations considering inexact computation of update steps: computational efficiency.

The gain in efficiency already becomes apparent as we take a look at the plot from Figure 4.2, where the number of required TNNMG iterations for computing the respective Proximal Newton trial update step is depicted. In particular, the Proximal Newton steps incorporate both accepted and declined iterates. Furthermore, we can recognize that the decrease of the forcing term η from the relative error criterion forces also the inexact version of our method to compute rather accurate solutions to the subproblems in the later stages of algorithm. This in particular enables the local superlinear convergence as we have verified in Theorem 4.1.4. In the globalization phase, however, it is easy to see that we spare many (apparently unnecessary) subproblem solver iterations and thus also save valuable computational time.

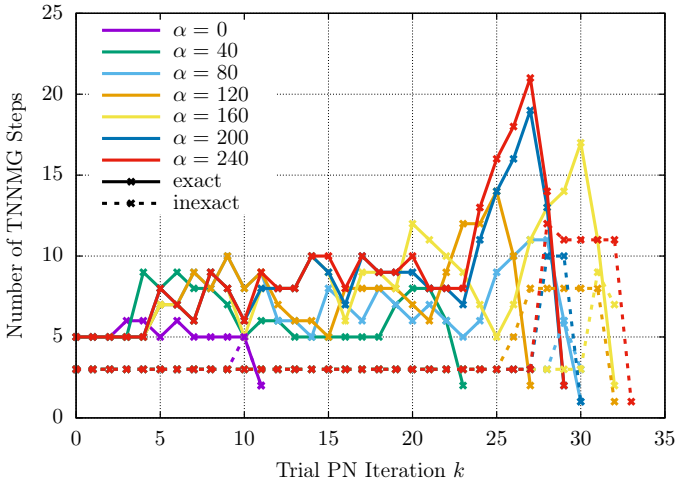


Figure 4.2: Number of TNNMG iterates required for update step computation in every trial Proximal Newton step for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$.

This reduction of required TNNMG steps can be ascribed to the inexactness criteria which we have introduced over the course of the current chapter. Even though it has been a central concern of ours to also provide efficient ways for the evaluation these prerequisites for inexact update steps, this still might negate our abovementioned gains in efficiency. In order to dispel this worry, we have recorded the essential data concerning overall algorithmic efficiency for both the exact and inexact variation of our method across all test scenarios in Table 4.1.

While the number of accepted ("Acc."), declined ("Decl."), and total Proximal Newton iterations required for finding the solution of the minimization problem overall are the same for both alternatives, both the number of total TNNMG iterations and wall-time needed for this endeavor reveals the gains in efficiency of the inexact method. In particular, the evaluation of inexactness criteria is included in the TNNMG wall-time share. The advantageous properties of the modified algorithm become more and more apparent as α and thereby the complexity of the underlying minimization problem increases. In order to give a clear illustration of the corresponding ratios and a comparison between the exact and inexact method, the latter information on wall-times is also depicted in Figure 4.3.

However, we have to note here that across all numerical tests here the determining factor for the total wall-time of the respective run is the time required by the assembler, i.e., the time it takes to compute gradients and hessians, and to from there establish the respective second

α	Variant	PN-Iterations			TNNMG-It.	Wall-Time in sec.		
		Acc.	Decl.	Total		TNNMG	Assembler	Total
0	Exact	11	3	14	60	13.99	89.02	117.40
	Inexact	11	3	14	37	9.37	88.81	112.67
40	Exact	15	9	24	147	30.74	109.68	166.94
	Inexact	15	9	24	72	15.43	109.52	152.12
80	Exact	19	12	31	214	44.79	139.86	218.99
	Inexact	19	12	31	96	20.67	139.45	195.11
120	Exact	17	11	28	211	43.97	124.78	201.52
	Inexact	18	15	33	124	26.49	134.69	198.99
160	Exact	19	14	33	271	56.59	139.11	232.25
	Inexact	19	14	33	109	23.41	140.70	201.61
200	Exact	17	13	30	254	52.99	131.44	217.46
	Inexact	18	13	31	105	22.60	138.89	196.66
240	Exact	18	12	30	268	56.01	131.81	220.97
	Inexact	18	16	34	141	30.27	142.25	211.08

Table 4.1: Comparative statistics for the exact and inexact variant of our Proximal Newton method for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$.

order problems which are then solved for the computation of Proximal Newton update steps. Furthermore, the wall-time shares of TNNMG and the assembler do not add up to the total time elapsed over one run of the algorithm since the latter additionally incorporates e.g. the evaluation of decrease criteria and update procedures of iterates and parameters.

Having in mind the goal of the introduction of inexactness, on the other hand, we can still declare this endeavor as a success. As far as the time for solving the step computation subproblems is concerned, we have spared 51.7% across the above numerical tests which is a significant improvement. In particular for problems where first and second order models can be computed explicitly without depending on automatic differentiation software, this gain in effectiveness is crucial.

Investigation of Inexactness Criteria

As mentioned beforehand, we also want to take a look at how the inexactness criteria affect the solution process of the step computation subproblems. To this end, we consider two aspects each of which covers one of our criteria based on exemplary computations of Proximal Newton steps: On the one hand, in order to investigate the relative error criterion (4.1.13), we compute every Proximal Newton step twice. Within the first computation, we neglect inexactness criteria which allows us to then compute the actual relative error E_{rel} of the TNNMG iterates in the second and actually inexact computation process. This makes it possible to compare the actual relative error to the estimate E_{est} which we use for easier evaluation, cf. (4.1.14).

As can be seen in the left-hand part of Table 4.2 and the plots in Figure 4.4 for representative trial step computations, both of these quantities stay within the same order of magnitude. This lets us infer that the employed triangle inequality for the deduction of (4.1.14) is surprisingly sharp in practice. Note that the estimated error E_{est} is not assigned within the first two TNNMG iterations since we have to take more of these into consideration in order to

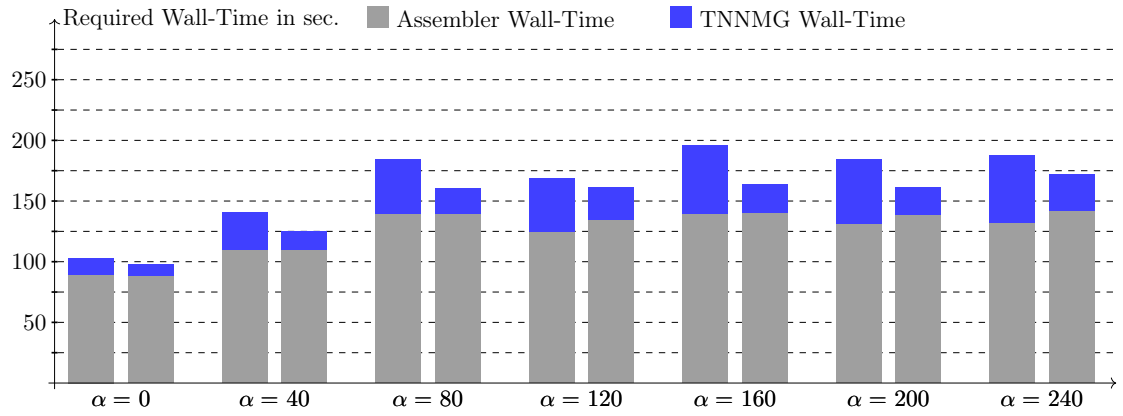


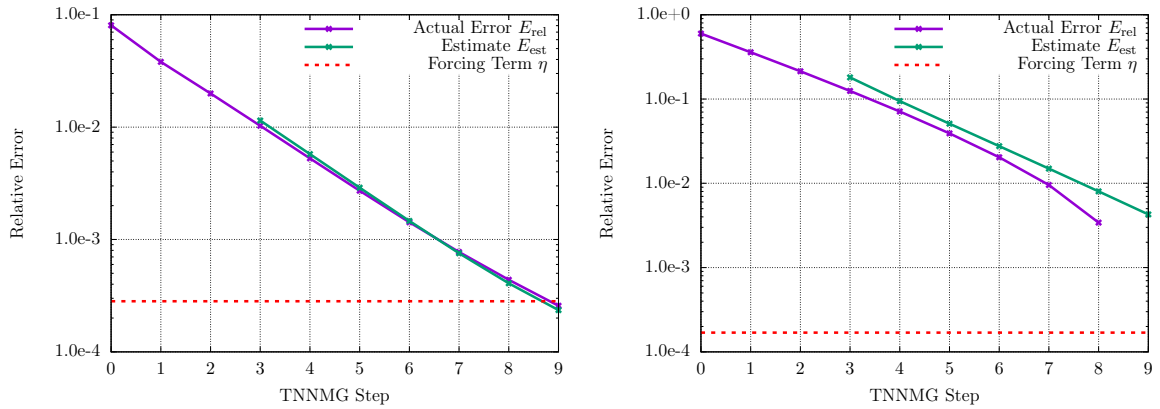
Figure 4.3: Assembler-, TNNMG-, and total wall-times required for algorithmic runs across the test series with $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$. The left bar represents the exact method and the right bar corresponds to the inexact variant.

obtain a valid estimate for multigrid convergence rates θ in (4.1.14). The respective column in Table 4.2 reveals that the estimated convergence rate then remains relatively constant over the minimization of the quadratic model which suggests it to be measured adequately by our procedure.

Furthermore, the graph for the computation of trial step $k = 29$ in Figure 4.4b shows that the forcing term in this case was so small that the relative error criterion (4.1.13) could not be met by iterates of the subproblem solver before the latter stopped computation due to the default criterion from TNNMG. Thus, the relative error to the (numerically) exact solution of the subproblem is zero for the last data point in the actual relative error which also explains why it is missing in the corresponding logarithmic plot. In particular, the exact computation of update steps close to optimal solutions is crucial for the local acceleration of our method as shown in Theorem 4.1.4. All in all, we conclude that the estimate which implicitly uses the convergence rate of our multigrid subproblem solver constitutes an adequate and easy-to-evaluate alternative to the actual relative error.

On the other hand, we also consider the subgradient inexactness criterion (4.1.25). As mentioned beforehand, we have introduced this criterion for globalization purposes with the intention that it would not interfere with the minimization process, especially in the local acceleration phase close to optimal solutions. In fact, we have noticed that throughout our tests the determining quantity for further solving the subproblem was the relative error estimate and not that $\tilde{\omega}$ from (4.1.27) was too large. For example, over the TNNMG-iterations of the Proximal Newton trial step considered in Figure 4.4a we had nearly constant $\tilde{\omega} \approx 8.5$, clearly remaining below our choice of $\tilde{\omega}_{\max} = 10^{10}$.

With these satisfying results concerning the improved efficiency of update step computation for our Proximal Newton method at hand, we can now turn our attention to other aspects which still show room for improvement in that regard.



(a) Proximal Newton trial step $k = 28$.

(b) Proximal Newton trial step $k = 29$.

Figure 4.4: Comparison of the relative error as required in (4.1.13) and its estimator from (4.1.14) within the computation of Proximal Newton trial steps $k = 28$ and $k = 29$ while minimizing (4.1.47) for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha = 200$.

i_k	E_{rel}	E_{est}	η	θ	$\tilde{\omega}$	$\tilde{\omega}_{\max}$
1	0.0805495	not assigned	0.000282111	1.03008e-06	8.71446	1e+10
2	0.0381125	not assigned	0.000282111	0.0487216	8.81993	1e+10
3	0.0199241	0.0111777	0.000282111	0.372641	8.79332	1e+10
4	0.0102817	0.0114414	0.000282111	0.533085	8.53583	1e+10
5	0.00527283	0.0057486	0.000282111	0.524131	8.91433	1e+10
6	0.00271465	0.00289124	0.000282111	0.517709	8.41266	1e+10
7	0.00142414	0.00146153	0.000282111	0.513961	8.8467	1e+10
8	0.00077351	0.000754427	0.000282111	0.514819	8.57349	1e+10
9	0.000438303	0.000407447	0.000282111	0.522954	8.31665	1e+10
10	0.000257378	0.0002354	0.000282111	0.539877	8.20661	1e+10

Table 4.2: Overview for inexactness criteria along TNNMG iterations i_k in Proximal Newton trial step $k = 28$ while minimizing (4.1.47) for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha = 200$. Listed are the actual relative error E_{rel}, its estimate E_{est}, the forcing term η, the estimated TNNMG convergence rate θ, the subgradient regularization parameter $\tilde{\omega}$, and its upper bound $\tilde{\omega}_{\max}$.

4.2 Choice of Parameters

As we have mentioned throughout the previous numerical investigations of the functionality of our Proximal Newton method both in its exact (cf. Section 3.2.7) and its inexact variant (cf. Section 4.1.6), the heuristic choices for the regularization parameter ω in (3.2.13) and the forcing terms η in (4.1.13) still show major room for improvement. While the previously employed strategies for finding adequate values of these algorithmic quantities have always been compatible with the corresponding results from convergence analysis of the underlying minimization procedure, they were not problem-specific and thus did not properly take advantage of structural peculiarities of the problem to be solved. In contrast to that, so-called *adaptive* approaches to the choice of algorithmic parameters make use of such properties of the objective functional and thereby hold out the prospect of significant improvements with regard to algorithmic efficiency of the considered method.

Section Outline

As has been the case for many algorithmic concepts which we have investigated beforehand, also the choice or parameters lets us draw a considerable amount of inspiration from the smooth case of Newton methods without an additional non-smooth part g in (3.2.1). The generalization of the corresponding ideas to our Proximal Newton scenario here together with algorithmic improvements will constitute the main part of the current section which is straightforwardly structured as follows: In Sections 4.2.1 and 4.2.2, respectively, we will discuss the demands on the regularization parameter ω and the forcing term η which arise from the convergence analysis conducted beforehand and then present as well as discuss adequate adaptive strategies for their choice within our algorithm. Section 4.2.3 then concerns the numerical comparison of these strategies both among themselves and with the heuristic approaches pursued in the previous numerical investigations.

4.2.1 Choice of the Regularization Parameter

Naturally, the demands on the choice of the regularization parameter $\omega \geq 0$ in (3.2.13) are intimately related to the results from convergence analysis deduced in the context of its use. Originally, ω has been introduced for two main reasons: Firstly, we have pursued the compensation of lacking convexity of our composite objective functional F from (3.2.1) insofar that the regularized second order decrease model $\lambda_{x,\omega}$ from (3.2.12) is strongly convex and thus update step computation allows for a unique solution. As far as this demand is concerned, ω always has to be chosen **sufficiently large** but adaptivity of the corresponding procedure of choice is rather peculiar. Our TNNMG subproblem solver can only detect whether the subproblem presented to it is convex or not. To which extent this non-convexity is present is not specifiable any further. For this reason, independent of the strategy which we use in order to choose ω, this problem is always solved by increasing ω multiplicating it with a fixed increase factor.

Global Convergence for Sufficient Regularization

The second demand which we impose on the choice of the regularization parameter is connected to its role within the globalization strategy of our algorithm. As we have pointed out during our deliberations considering the development of sufficient decrease criteria in Section 3.2.6, the globalization strategy, which is mostly governed by a corresponding parameter, has to make

sure that the sufficient decrease criterion yielding global convergence results is satisfiable. For the decrease parameter $\gamma \in]0,1]$, in our framework, the sufficient decrease criterion has been given in (3.2.15) by

$$F\big(x+\Delta s(\omega)\big) - F(x) \leq \gamma \lambda_{x,\omega}\big(\Delta s(\omega)\big) \tag{4.2.1}$$

in the general case with a modification for numerical stability considered in Section 4.1.5 via

$$\big[f'\big(x+\Delta s(\omega)\big) - f'(x) - \big(H_x + \omega\mathfrak{R}\big)\big(\Delta s(\omega)\big)\big]\Delta s(\omega) \leq -\frac{1+\gamma}{2}\omega\big\|\Delta s(\omega)\big\|_X^2 . \tag{4.2.2}$$

For both of these criteria, we have verified that they are satisfied once the regularization parameter ω is chosen **sufficiently large**, cf. Lemmas 3.2.12 and 3.2.30 or Lemmas 4.1.5 and 4.1.11, respectively.

Even though this constitutes the second demand on ω itself, the corresponding demand on its choice has to be formulated in a slightly different way. In view of globalization, the strategy of finding an adequate value for the regularization parameter has to be able to adaptively identify whether the respective sufficient decrease criterion is satisfied or not. If it comes to the conclusion that (4.2.1) or (4.2.2) is violated, also the information to which extent this is the case is of importance in order to then **increase $\boldsymbol{\omega}$ accordingly**. Thus, we can see that the update of the regularization parameter has to be somewhat tailored to the sufficient decrease criterion for adaptive approaches.

Local Acceleration for Vanishing Regularization

While we have seen that the second demand on the choice of the regularization parameter has rather emphasized the restrictive nature of sufficient decrease criteria far away from optimal solutions of the underlying minimization problem for the sake of globalization, the third and last demand concerns the topic of releasing the regularization of the model in order to achieve fast local convergence. As has been verified for both the original sufficient decrease criterion (4.2.1) and its numerically robust variant (4.2.2) in Propositions 3.2.28, 3.2.32, and 4.1.10, 4.1.13, respectively, the theory behind our Proximal Newton method allows for the arbitrarily small choice of the regularization parameter close to optimal solutions of (3.2.1) at which the objective functional exhibits sufficient convexity and semi-smoothness properties.

This "permission" to use a (nearly) unregularized model close to optimal solutions thus also has to be taken advantage of by our adaptive choice in order to benefit from the local superlinear convergence established both in the exact and inexact case in Theorems 3.2.8 and 4.1.4, respectively. In order to quickly identify whether the region of theoretical local convergence has already been reached within the globalization phase of our algorithm, our procedure thus also here has to be able to quantify to which extent the sufficient decrease criterion is satisfied. Then, it has to **reduce $\boldsymbol{\omega}$ according to this information** in order to relax possible "overregularization" of the problem at the current iterate. In conclusion, we can say that the above arguments again emphasize that the update of ω has to be somewhat tailored to the sufficient decrease criterion and they imply that our regularization procedure has to create a **sequence $(\omega_k)_{k\in\mathbb{N}}$ tending to zero** if the underlying problem structure allows so.

We can summarize the aspects from above as follows: Firstly, the **well-definedness** of (3.2.13), i.e., $\omega > -\big(\kappa_1(x)+\kappa_2\big)$; secondly, the **sufficient decrease** criterion for which ω also

has to be sufficiently large and lastly the **convergence to zero** as the sequence of iterates approaches an optimal solution. All in all, it is desirable to choose the regularization parameter as large as necessary but at the same time as small as possible.

Reconsidering the Decrease Ratio Functions

While the requirement of well-definedness of update steps still has to be tackled rather manually once the subproblem solver discovers non-convexity of the step computation minimization problem (3.2.13), we will recognize that the remaining two can be tied to the decrease ratio function

$$\gamma\colon X \times [0,\infty[\to \mathbb{R}\,, \quad \gamma(x,\omega) \coloneqq \frac{F\big(x_+(\omega)\big) - F(x)}{\lambda_{x,\omega}\big(\Delta s(\omega)\big)}\,. \tag{4.2.3}$$

This mapping has made its first appearance in the proof of Proposition 3.2.28 where we in particular have elaborated on its convergence to (something greater equal than) one, cf. (3.2.39). For the numerically stable formulation with (4.2.2), we have accordingly defined the alternative decrease ratio function

$$\tilde{\gamma}\colon X \times [0,\omega[\to \mathbb{R}\,, \quad \tilde{\gamma}(x,\omega) \coloneqq \frac{\big[f'\big(x+\Delta s(\omega)\big) - f'(x) - (H_x + \omega\mathfrak{R})\Delta s(\omega)\big]\Delta s(\omega)}{-\omega\big\|\Delta s(\omega)\big\|_X^2} \tag{4.2.4}$$

in (3.2.43) and also there explained its convergence to some value greater equal than one which will be of use for us later on.

The Heuristic Strategy

In order to later on refer to it for the comparison to our newly defined adaptive strategies, let us first give a formulation of the heuristic procedure pursued in Sections 3.2.7 and 4.1.6. There, the regularization parameter has been adjusted solely based on the success of the current (and preceding) update steps according to the corresponding sufficient decrease criterion. We denote this choice (in dependence of the previously used regularization parameter ω) by

$$[\omega]_1(\omega) \coloneqq \begin{cases} \left(\frac{1}{2}\right)^n \omega, & \text{if the } n \geq 1 \text{ preceding updates were successful} \\ 2\omega, & \text{if the current update was not admissible} \end{cases}, \tag{4.2.5}$$

where admissibility depends on either (4.2.1) or (4.2.2). As mentioned beforehand, this strategy does not take into account the structure of the underlying problem. On a different note, it is an easy and fitting choice for the convergence theory from Sections 3.2 and 4.1 since it trivially increases the regularization parameter in the case of unaccepted update steps and makes ω tend to zero as we approach optimal solutions since close to these arbitrarily small values of ω are permissible due to the respective transition results. However, in the following we will provide two suitable adaptive alternatives for the choice of ω: The first one is directly tied to the underlying sufficient decrease criterion and also takes into account to what extent it has been violated by the current update step.

The Controller Strategy

Its derivation takes some inspiration from control theory. To be exact, it originates from the adaptive control of single step methods in numerics for ordinary differential equations to obtain a guess for the next time step size τ_{j+1} based on the current step size τ_j, cf. [21, Chapter 5.2].

There, similar problems have to be faced: On the one hand, the adaptively determined step sizes for initial value problems have to be as large as possible in order to minimize necessary computational effort for finding the solution. On the other hand, step sizes also should remain sufficiently small such that errors with respect to the exact solution do not become too large. In our case, regularization parameters can be understood as "the inverse" of step sizes in the above sense: As we have quantified in the estimate deduced in Lemmas 3.2.26 and 4.2.1, while large regularization parameters cause update steps to be small, small values of ω lead to less damping of the respective step also in norm.

Let us now first take a look at how the above demands are realized within control theory for ODEs in [21]. To this end, we consider the so-called *adaptive base algorithm* which can be summarized as follows: At first, some *local error quantity* $|[\varepsilon_{j+1}]|$ is determined which should be within the same order of magnitude as the *prescribed tolerance* TOL for such an error. Obviously, $|[\varepsilon_{j+1}]|$ should not be too large (reliability) but also not too small (efficiency) since small values for this error quantity are assumed to be only a result of unreasonably high computational effort opposed to random influences or some lucky choice. From there, rather ODE-specific arguments lead to the formula

$$\tau_{j+1} \coloneqq \sqrt[p+1]{\frac{\rho \cdot \mathrm{TOL}}{|[\varepsilon_{j+1}]|}}\, \tau_j \tag{4.2.6}$$

for the next step size τ_{j+1} where $\rho < 1$ denotes some safety factor and $p \in \mathbb{N}$ is the so-called *order of the employed method*. We will later on further specify the latter quantity and deduce the corresponding value for our (inexact) Proximal Newton method. Now, our goal is to translate the structure of an estimator for step sizes from (4.2.6) to an adaptive strategy for choosing the regularization parameter ω in our scenario.

The key features of the above formula can be conceived as the error estimator $|[\varepsilon_{j+1}]|$ and the TOL-term which means that we will first derive corresponding expressions for these quantities. For this reason, let us simplify the notation within our sufficient decrease criteria (4.2.1) and (4.2.2): In the general case, for some $x \in X$ and $\omega \in [0, \infty[$, we thus define the *actual reduction* $a_{\mathrm{red}}(x, \omega)$ and the *predicted reduction* $p_{\mathrm{red}}(x, \omega)$ via

$$a_{\mathrm{red}}(x,\omega) \coloneqq F\big(x + \Delta s(\omega)\big) - F(x) \quad \text{and} \quad p_{\mathrm{red}}(x,\omega) \coloneqq \lambda_{x,\omega}\big(\Delta s(\omega)\big)$$

where $\Delta s(\omega)$ denotes some inexact update step within our Proximal Newton method from Algorithm 12 computed at x with regularization parameter ω. As a consequence, the corresponding criterion (4.2.1) takes the form $a_{\mathrm{red}}(x,\omega) \leq \gamma p_{\mathrm{red}}(x,\omega)$. We carry over these notational principles also to the numerically stable case by defining

$$\begin{aligned} &\tilde{a}_{\mathrm{red}}(x,\omega) \coloneqq \big[f'\big(x + \Delta s(\omega)\big) - f'(x) - (H_x + \omega\mathfrak{R})\Delta s(\omega)\big]\Delta s(\omega)\,, \\ &\tilde{p}_{\mathrm{red}}(x,\omega) \coloneqq -\omega\big\|\Delta s(\omega)\big\|_X^2 \quad \text{and} \quad \tilde{\gamma} \coloneqq \frac{1+\gamma}{2} \in \,]\frac{1}{2}, 1]\,. \end{aligned} \tag{4.2.7}$$

Obviously, the specific interpretations of actual and predicted decrease break down in the numerically stable case but the respective sides of the corresponding sufficient decrease criterion will play the same role both within global convergence theory and the choice of the regularization parameter. With these definitions, the alternative sufficient decrease criterion (4.2.2) takes the similar form $\tilde{a}_{\mathrm{red}} \leq \tilde{\gamma}\tilde{p}_{\mathrm{red}}$.

With these reformulations in mind, we can now tackle the transformation of the tolerance and relative error term in (4.2.6). To that end, we consider the border case where the respective sufficient decrease criterion is "tightly" satisfied. We do not use the tilde notation here, but the same arguments can be made for the respective quantities in the numerically stable case. We perceive that

$$\begin{aligned} a_{\mathrm{red}}(x,\omega) \approx \gamma p_{\mathrm{red}}(x,\omega) \quad &\Leftrightarrow \quad a_{\mathrm{red}}(x,\omega) - p_{\mathrm{red}}(x,\omega) \approx (\gamma-1)p_{\mathrm{red}}(x,\omega) \\ &\Leftrightarrow \quad \frac{p_{\mathrm{red}}(x,\omega) - a_{\mathrm{red}}(x,\omega)}{p_{\mathrm{red}}(x,\omega)} \approx 1-\gamma \end{aligned}$$

holds here. The left-hand side of the latter approximate equality can be understood as a relative error estimate of the quadratic model in comparison to the actual non-linearity of F which is why we will also interpret the right-hand side $(1-\gamma)$ as the corresponding TOL-term. Here, we apply the same logic as for step sizes regarding efficiency and reliability.

In view of regularization parameters as inverse step sizes we can now carry over the formula from (4.2.6) to our scenario via

$$[\bar{\omega}]_2(\omega) \coloneqq \left(\sqrt[\theta]{\frac{\rho\cdot(1-\gamma)}{\left|\frac{p_{\mathrm{red}}(x,\omega)-a_{\mathrm{red}}(x,\omega)}{p_{\mathrm{red}}(x,\omega)}\right|}} \right)^{-1} \omega = \sqrt[\theta]{\frac{\left|1-\frac{a_{\mathrm{red}}(x,\omega)}{p_{\mathrm{red}}(x,\omega)}\right|}{\rho\cdot(1-\gamma)}}\,\omega \tag{4.2.8}$$

where we are still free to choose the safety factor $\rho < 1$ and the root order $\theta \geq 1$ adequately. Let us add some remarks considering the latter: In the adaptive base algorithm (4.2.6), we noted that $\theta = p+1$ for p the order of the method constitutes an admissible choice for this quantity. With the aid of more sophisticated control theory for PID-controllers, however, it is possible to see that we should choose $\theta > \frac{p+1}{2}$ close to that lower bound, cf. [21, Section 5.2.2].

In particular, the "order" p of our method is not to be confused with the order of the model employed for update step computation. According to its definition, p measures the linearized discrepancy between the regularized model and the actual non-linearity of F in dependence of the regularization parameter ω. In order to determine this quantity, we first prove the following lemma which states that the update step norm actually scales directly with the regularization parameter:

Lemma 4.2.1: Influence of Regularization on Update Step Norms

Let $\Delta x(\omega)$ be an update step exactly computed at $x \in X$ according to (3.2.13) for some $\omega > -\big(\kappa_1(x)+\kappa_2\big)$. Then, for any $\mu \in \partial_F g(x)$, the following estimate holds:

$$\big\|\Delta x(\omega)\big\|_X \leq \frac{\big\|f'(x)+\mu\big\|_{X^*}}{\omega+\kappa_1(x)+\kappa_2}\,.$$

Proof. First, we will derive an auxiliary estimate considering Fréchet-subdifferential elements $\mu \in \partial_F g(x)$. By the convexity assumption (A4), $\bar{g} \coloneqq g + \frac{\kappa_2}{2}\|\cdot\|_X^2$ is convex and thus, at every $y \in X$, subdifferential elements $\bar{\mu} \in \partial\bar{g}(y)$ satisfy $\bar{g}(z) \geq \bar{g}(y) + \bar{\mu}(z-y)$ for any $z \in X$.

This in turn yields that for every $y \in X$ and $\mu \in \partial_F g(y)$ the inequality

$$g(z) \geq g(y) + \mu(z-y) + \frac{\kappa_2}{2}\big\|z-y\big\|_X^2 \tag{4.2.9}$$

holds for any $z \in X$.

In the following computation, we take advantage of (4.2.9) with $y = x_+(\omega)$, $z = x$ and $\mu_+ \in \partial_F g\big(x_+(\omega)\big)$ from (3.2.41) as well as $y = x$, $z = x_+(\omega)$ and arbitrary $\mu \in \partial_F g(x)$:

$$\begin{aligned}\big(\omega + \kappa_1(x) + \frac{\kappa_2}{2}\big)\big\|\Delta x(\omega)\big\|_X^2 &\leq (\omega\mathfrak{R} + H_x)\big(\Delta x(\omega)\big)^2 + \mu_+\Delta x(\omega) + g(x) - g\big(x_+(\omega)\big)\\ &= -f'(x)\Delta x(\omega) + g(x) - g\big(x_+(\omega)\big)\\ &\leq -\big(f'(x)+\mu\big)\Delta x(\omega) - \frac{\kappa_2}{2}\big\|\Delta x(\omega)\big\|_X^2\end{aligned}$$

Equivalently, we obtain

$$\big(\omega + \kappa_1(x) + \kappa_2\big)\big\|\Delta x(\omega)\big\|_X^2 \leq -\big(f'(x)+\mu\big)\Delta x(\omega)$$

which directly implies the asserted estimate for the norm of exactly computed update steps. □

Remark. *This inequality can be interpreted in two ways in order to conclude small update step norms close to optimal solutions. On the one hand, we have verified in Proposition 3.2.14 that the numerator of the upper bound converges to zero over the course of our algorithm. On the other hand, the above result also quantifies the indirect proportionality between length of update steps and the regularization parameter.*

Let us now continue with the determination of the order of our method. For the sake of simplicity, we now consider a rather smooth scenario with

$$\big|f(x+\delta x) - \big(f(x) + f'(x)\delta x + \frac{1}{2}H_x(\delta x)^2\big)\big| = O\big(\big\|\delta x\big\|_X^3\big)$$

in the limit of $\delta x \to 0$ which in the exact case provides us with

$$\begin{aligned}\big|\frac{p_{\mathrm{red}}(x,\omega) - a_{\mathrm{red}}(x,\omega)}{p_{\mathrm{red}}(x,\omega)}\big| &= \big|1 - \frac{\lambda_{x,\omega}\big(\Delta x(\omega)\big) - \frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 + O\big(\big\|\Delta x(\omega)\big\|_X^3\big)}{\lambda_{x,\omega}\big(\Delta x(\omega)\big)}\big|\\ &\leq \frac{\frac{\omega}{2}\big\|\Delta x(\omega)\big\|_X^2 + O\big(\big\|\Delta x(\omega)\big\|_X^3\big)}{\big|\lambda_{x,\omega}\big(\Delta x(\omega)\big)\big|}\\ &\leq \frac{\omega}{\omega + \kappa_1(x) + \kappa_2} + \frac{2}{\omega + \kappa_1(x) + \kappa_2}\xi\big(\Delta x(\omega)\big)\end{aligned}$$

for the relative error estimate. For the latter term here, $\xi\big(\Delta x(\omega)\big) = O\big(\big\|\Delta x(\omega)\big\|_X\big)$, we can then use Lemma 4.2.1 in order to identify the relative error estimate in dependence of the regularization parameter as

$$|\varepsilon(\omega)| \coloneqq \big|\frac{\omega}{\omega + \kappa(x)} + \frac{2C}{(\omega+\kappa(x))^2}\big|$$

for some constant $C > 0$ and $\kappa(x) \coloneqq \kappa_1(x) + \kappa_2$. The case of $\kappa(x) = 0$ can be viewed as the "most instable" one here and the linearization around some $\omega_0 \geq 0$ then takes the form

$$|\varepsilon'(\omega_0)| = \big|\frac{1}{\omega_0 + \kappa(x)} - \frac{\omega_0}{\big(\omega_0 + \kappa(x)\big)^2} - \frac{4C}{\big(\omega_0 + \kappa(x)\big)^3}\big| = \frac{4C}{\omega_0^3}\,.$$

Now, we can identify the order of our regularization method as $p = 2$ and thus obtain $\theta > \frac{3}{2}$ in (4.2.8) close to that lower bound as a suggestion from control theory.

Even though well motivated, the update strategy for ω defined via (4.2.8) still involves a major flaw in the light of our Proximal Newton methods. Due to the symmetry inherent to the absolute value term for the error estimator, $a_{\text{red}} < (1+\gamma)p_{\text{red}}$ leads to a prefactor larger than one and thus to an increase within the regularization parameter. This fact is illustrated in Figure 4.5a. Such behavior is counter-intuitive since large (negative) actual descent signifies a desirable property of the corresponding update step and should lead to a decrease of ω. For this reason, we adapt the update strategy from (4.2.8) as follows:

For comparatively bad updates, i.e., $a_{\text{red}} \geq p_{\text{red}}$, the above choice works trouble-free which is why we hold on to it there. For $a_{\text{red}} < p_{\text{red}}$ however, we want the prefactor function to stay below one while still mimicking the "absolute-value-like" behavior of the error estimator. Consequently, we define the *radicand function* depending on parameters $p < 0$, $\rho \in]0,1]$ and $\gamma \in]0,1]$ as follows

$$\text{Rad}_{\rho,\gamma,p} \colon \mathbb{R} \to [0,\infty[\quad , \quad \text{Rad}_{\rho,\gamma,p}(a) \coloneqq \begin{cases} \dfrac{p-a}{p-\big(1+\rho(1-\gamma)\big)a} & , \text{ for } a < p\,, \\ \dfrac{1-\frac{a}{p}}{\rho(1-\gamma)} & , \text{ for } a \geq p\,. \end{cases} \tag{4.2.10}$$

The desirable (and obvious) properties of the radicand function can be summarized as follows:

Lemma 4.2.2: Properties of the Radicand Function

The radicand function as defined in (4.2.10) is ...

(i) ... continuous on $\mathbb{R}$ with $\lim\limits_{a\to-\infty} \text{Rad}_{\rho,\gamma,p}(a) = \frac{1}{1+\rho(1-\gamma)}$, $\text{Rad}_{\rho,\gamma,p}(p) = 0$ as well as

$$\text{Rad}_{\rho,\gamma,p}(a) > \frac{1}{\rho} \text{ for } a > \gamma p \quad \text{and} \quad \text{Rad}_{\rho,\gamma,p}(a) < 1 \text{ for } a \leq p\,.$$

(ii) ... continuously differentiable on $\mathbb{R} \setminus \{p\}$ with one-sided derivatives

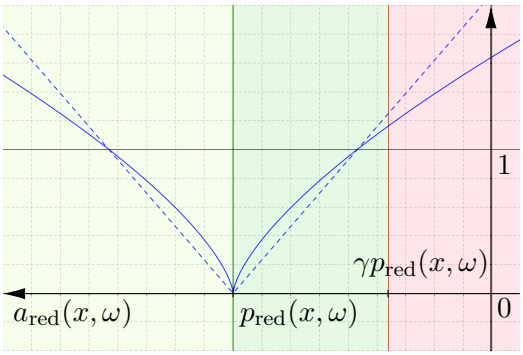
$$-\text{Rad}'_{\rho,\gamma,p}(p;-1) = \frac{1}{\rho(1-\gamma)p} = \text{Rad}'_{\rho,\gamma,p}(p;1)\,.$$

With the radicand function and its properties at hand we can introduce the modified version of the controller strategy (4.2.8) by

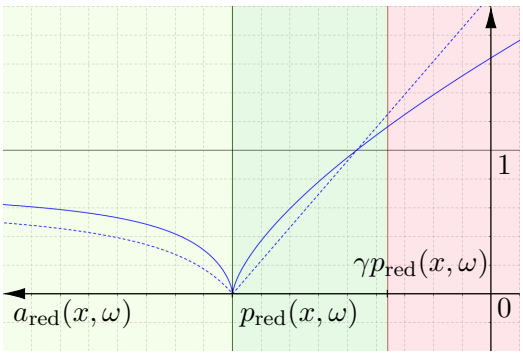
$$[\omega]_2(\omega) \coloneqq \sqrt[\theta]{\text{Rad}_{\rho,\gamma,p_{\text{red}}(x,\omega)}\big(a_{\text{red}}(x,\omega)\big)}\,\omega \tag{4.2.11}$$

with $\text{Rad}_{\rho,\gamma,p} \colon \mathbb{R} \to [0,\infty[$ from (4.2.10), $\theta > \frac{3}{2}$ close to that lower bound, $\rho \in]0,1]$ a safety factor, and $\gamma \in]0,1]$ the sufficient decrease parameter from (4.2.1). The advantageous properties of the prefactor function from (4.2.11) defined via the radicand function from (4.2.10) in contrast to the absolute value formulation are illustrated in Figure 4.5b.

Let us shortly remark on the safety factor $\rho < 1$ and its importance for the well-behavedness of our procedure: It does not only bring the choice (4.2.11) in accordance with its motivational formula from control theory (4.2.6) but also provides a minimal relative increase of the regularization parameter in case of the update step failing the sufficient decrease criterion. This



(a) Absolute Value Base (dashed line) and Prefactor (full line) Function.



(b) Radicand Base (dashed line) and Prefactor (full line) Function.

Figure 4.5: Base and prefactor functions for different approaches of the controller strategy: For some fixed predicted reduction $p_{\mathrm{red}}(x,\omega)$ the corresponding function values are illustrated in dependence on the actual reduction $a_{\mathrm{red}}(x,\omega)$. While update steps are rejected in the red area, the remaining two areas signify admissibility of the step. Here, we refer to the expressions under the respective θ-roots as *Base* and to the ones modified by said root as *Prefactor Functions*.

helps us find an update step yielding sufficient decrease within few iterations in contrast to the case of $\rho = 1$ where one might get stuck with non-sufficient decrease in practice. Within the span of $]0,1]$, high values of ρ promote small choices of the regularization parameter and thus a rather risk-taking procedure whereas small ones lead to relatively large ω and thereby a more conservative strategy for computing trial update steps. This can also be retraced in Figure 4.5 where in both illustrations a safety factor slightly below one has been used. For this reason, the corresponding base functions have the value $1/\rho > 1$ at the border case of $a_{\mathrm{red}} = \gamma p_{\mathrm{red}}$. In particular, this implies an increase in ω even if the corresponding update step is only just admissible. In this light, ρ can also be viewed as an "inverse safety factor" but we will stick to its original designation anyhow.

The controller strategy for the regularization parameter as formulated in (4.2.11) has been developed for the sufficient decrease criterion from (4.2.1) and has to be slightly adapted in case the current iterate is located close to an optimal solution and thus our method demands for a numerically more stable computation scheme. As mentioned beforehand, our adapted nomenclature from (4.2.7) turns out to be convenient for this endeavor and allows us to simply define

$$[\tilde{\omega}]_2(\omega) \coloneqq \sqrt[\theta]{\mathrm{Rad}_{\rho,\tilde{\gamma},\tilde{p}_{\mathrm{red}}(x,\omega)}\big(\tilde{a}_{\mathrm{red}}(x,\omega)\big)}\,\omega \tag{4.2.12}$$

with again $\mathrm{Rad}_{\rho,\gamma,p}\colon \mathbb{R} \to [0,\infty[$ from (4.2.10), $\theta > \frac{3}{2}$ close to that lower bound, $\rho \in]0,1]$ the safety factor, and $\tilde{\gamma} = \frac{1+\gamma}{2}$ with $\gamma \in]0,1]$ the sufficient decrease parameter from (4.2.2).

Now, we are in a position to elaborate on how the above modified controller strategies implement our requirements for the choice of regularization parameters:

Proposition 4.2.3: Regularization Increase for the Controller Strategy

Consider $\omega \geq 0$ such that the corresponding update step $\Delta s(\omega)$ is not admissible for sufficient decrease according to (4.2.1) or (4.2.2), respectively.

Then, we have $[\omega]_2(\omega) > \sqrt[\theta]{\frac{1}{\rho}}\,\omega > \omega$ for (4.2.11) and analogously for (4.2.12).

Proof. For unaccepted update steps we have $a_{\mathrm{red}} > \gamma p_{\mathrm{red}}$, thereby $\mathrm{Rad}_{\rho,\gamma,p_{\mathrm{red}}}(a_{\mathrm{red}}) > 1/\rho$ via Lemma 4.2.2 which implies the assertion. The same behavior can be observed in the numerically stable case using (4.2.2) and (4.2.12). □

The remaining important property of our controller strategy for ω considers its convergence behavior close to optimal solutions of (3.2.1). While for the heuristic choice the admissibilty of arbitrarily small values of the regularization parameter yielded its convergence to zero, we have to incorporate more deliberations to this result here.

In particular, it will be important that the second order bilinear forms H_x satisfy the bound (3.2.36) which is given by

$$\big(H_{x_+(\omega)} - H_x\big)\big(x_+(\omega) - x_*\big)^2 = o\big(\big\|x - x_*\big\|_X^2\big) \text{ for } x \to x_* \tag{4.2.13}$$

and has already been demanded for the transition results in Propositions 3.2.28 and 4.1.10. As we have mentioned there, it is easy to see that the bound holds if either we have uniform boundedness of the second order bilinear forms together with superlinear convergence of the iterates or we have continuity of the mapping $x \mapsto H_x$ together with mere convergence of the iterates to x_*. The latter continuity assumption has been required in the proofs of the transition results for the numerically robust formulation in Propositions 3.2.32 and 4.1.13.

In our case here, however, we will use this estimate in order to show properties of the adaptive choice for both the regularization parameters ω_k above and forcing terms η_k (cf. Section 4.2.2) which is why a priori we can not assume superlinear convergence to hold since local convergence rates depend on the choice of these algorithmic quantities, cf. Theorem 4.1.4.

Ironically, generic and non-problem-specific null-sequence choices for ω and η in theory would thus get along with mere uniform boundedness of the H_x but adaptive strategies have the prospect of better performance for different application scenarios.

Proposition 4.2.4: Limit Behavior of the Controller Strategy

Suppose that all of the assumptions required for the respective transition result from either Proposition 4.1.10 for the original case or Proposition 4.1.13 for the numerically stable formulation hold.

Then, the values of the regularization parameter estimator defined via (4.2.11) or (4.2.12), respectively, converge to zero along the sequence of iterates of the inexact Proximal Newton method.

Proof. Since the procedure will always provide update steps satisfying the sufficient decrease criterion (4.2.1), global convergence of the ensuing method to an optimal solution is ensured. The decrease ratio functions from (4.2.3) and (4.2.4) in our new notation take the form

$$\gamma(x,\omega) = \frac{a_{\mathrm{red}}(x,\omega)}{p_{\mathrm{red}}(x,\omega)} \quad \text{and} \quad \tilde{\gamma}(x,\omega) = \frac{\tilde{a}_{\mathrm{red}}(x,\omega)}{\tilde{p}_{\mathrm{red}}(x,\omega)}\,,$$

respectively. Under the assumptions from the corresponding transition result, both of these converge to some value greater equal than one as we approach the solution.

By the properties of the radicand function from (4.2.10), this yields a prefactor of the regularization parameter which is (uniformly) smaller than the constant $1/\big(1+\rho(1-\gamma)\big) < 1$. This in turn allows us to conclude the convergence of ω to zero. □

While these results provide us at least with a theoretical justification for our construction, we will extensively study the behavior of the controller strategy from (4.2.11) and (4.2.12) within our numerical tests in Section 4.2.3. Additionally, we will compare the performance of the resulting Proximal Newton method with the one stemming from the other strategies for choosing the regularization parameter.

However, it still remains to be clarified exactly in which scenarios we will use the original update formula from (4.2.11) and in which scenarios we will resort to the numerically stable formulation of (4.2.12). In short, this decision depends on the sufficient decrease criterion which we have used in order to either admit or reject the preceding trial update step. Far away from optimal solutions of (3.2.1) we exclusively use the original formulation from (4.2.1) and thus also exclusively update the regularization parameter using (4.2.11). Once our sequence of iterates comes sufficiently close to optimal solutions, the following algorithmic scheme determines the computation of ω: As described in Algorithm 12, we first check (4.2.1) and admit the trial update step if it is fulfilled. In this case, we also use (4.2.11) for the update of ω. If (4.2.1) fails, however, we use (4.2.2) in order to check the decrease properties of the step also with respect to numerical cancellation. Regardless of the outcome, we then use the numerically stable formulation (4.2.12) in order to determine the new regularization parameter. The above strategy can be also retraced in the algorithmic depiction of Figure 4.11.

The Remainder Term Strategy

For the third and last strategy for choosing ω, we will step away from direct adjustments of the sufficient decrease criterion and augment an idea from [115] to our scenario of Proximal Newton methods. Considering the modified second order decrease model

$$\lambda_{x,\omega}(\delta x) = f'(x)\delta x + \frac{1}{2}H_x(\delta x)^2 + \frac{\omega}{2}\big\|\delta x\big\|_X^2 + g(x+\delta x) - g(x)$$

which we also use for step computation in (3.2.13), we recognize that the norm term from our regularization strategy should optimally replicate an actual remainder term for the nonlinearity of the smooth part f, at least in direction of the current update step. With this motivation in mind, we can simply demand the identity

$$\lambda_{x,\omega}\big(\Delta s(\omega)\big) \overset{!}{=} F\big(x+\Delta s(\omega)\big) - F(x)$$

and rewrite it as a requirement for the regularization parameter ω. This will only give us correspondence of norm regularization with the actual remainder term at the previous iterate in direction of the current one but hopefully give a good approximation also for the new optimal value of ω from this perspective.

Consequently, let $\Delta s(\omega)$ again denote an inexact update step satisfying (4.1.13) as well as (4.1.25) and consider an estimator of the form

$$[\omega]_3(\omega) \coloneqq \frac{2}{\rho\big\|\Delta s(\omega)\big\|_X^2}\Big|f\big(x+\Delta s(\omega)\big) - f(x) - f'(x)\Delta s(\omega) - \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2\Big| \quad (4.2.14)$$

where as before $\rho \in]0,1]$ denotes a safety factor with the same interpretation. While for the second strategy the sufficient decrease criterion (4.2.1) and control theory motivated the choice, here the remainder term intuition provides plausibility even though we cannot really speak of an actual Taylor remainder term due to lacking second order differentiability of f.

Also the strategy presented in (4.2.14) has to be adapted to the numerically robust formulation. For that reason, we use the following alternative choice close to optimal solutions where (4.2.2) determines admissibility of update steps:

$$[\tilde{\omega}]_3(\omega) \coloneqq \frac{2}{(1-\gamma)\rho\big\|\Delta s(\omega)\big\|_X^2}\big|\big[f'\big(x+\Delta s(\omega)\big) - f'(x) - H_x\big(\Delta s(\omega)\big)\big]\Delta s(\omega)\big|\,. \quad (4.2.15)$$

Here, instead of augmenting the second order model of f with an adequate remainder term as in (4.2.14), the first order model of f' has to be extended sufficiently. The prefactor fraction term stems from rather technical arguments which become apparent over the course of the latter proofs for properties of this choice.

As required before, recomputing the quantities from above after the corresponding sufficient decrease criterion has failed for some trial update step should increase the value of the estimator. This property of the above definitions can be retraced as follows:

Proposition 4.2.5: Regularization Increase for the Remainder Term Strategy

Consider $\omega \geq 0$ such that the respective update step $\Delta s(\omega)$ is not admissible for sufficient decrease according to (4.2.1) or (4.2.2), respectively.

Then, we have $[\omega]_3(\omega) > \frac{1}{\rho}\omega$ for the update from (4.2.14) and analogously for (4.2.15).

Proof. Let us start with the original sufficient decrease criterion: Since $\Delta s(\omega)$ does not yield sufficient decrease according to (4.2.1), we obtain the estimate

$$\begin{aligned}
[\omega]_3(\omega) &= \frac{2}{\rho\big\|\Delta s(\omega)\big\|_X^2}\big|f\big(x+\Delta s(\omega)\big) - f(x) - f'(x)\Delta s(\omega) - \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2\big| \\
&= \frac{2}{\rho\big\|\Delta s(\omega)\big\|_X^2}\big|F\big(x+\Delta s(\omega)\big) - F(x) - \lambda_{x,\omega}\big(\Delta s(\omega)\big) + \frac{\omega}{2}\big\|\Delta s(\omega)\big\|_X^2\big| \\
&> \frac{2}{\rho\big\|\Delta s(\omega)\big\|_X^2}\big|\frac{\omega}{2}\big\|\Delta s(\omega)\big\|_X^2 - (1-\gamma)\lambda_{x,\omega}\big(\Delta s(\omega)\big)\big|\,.
\end{aligned}$$

For the λ-term within the above estimate, we remember from (4.1.29) that inexactly computed update steps with (4.1.13) and (4.1.25) satisfy

$$\lambda_{x,\omega}\big(\Delta s(\omega)\big) \leq -\frac{\gamma}{2}\big\|\Delta x(\tilde{\omega}_{\max}+\omega+1)\big\|_X^2 \leq 0\,.$$

Inserting this finding into the above estimate directly completes the proof in the generic case.

The corresponding result for the numerically stable formulation (4.2.15) can be verified in a similarly easy fashion. To this end, we consider an inexact update step which does not

satisfy (4.2.2) in the form (3.2.42) and directly obtain the desired result via

$$\begin{aligned}[\widetilde{\omega}]_3(\omega) &= \frac{2}{(1-\gamma)\rho\big\|\Delta s(\omega)\big\|_X^2}\big|\big[f'\big(x+\Delta s(\omega)\big) - f'(x) - H_x\big(\Delta s(\omega)\big)\big]\Delta s(\omega)\big| \\ &> \frac{2}{(1-\gamma)\rho\big\|\Delta s(\omega)\big\|_X^2}\frac{1-\gamma}{2}\omega\big\|\Delta s(\omega)\big\|_X^2 = \frac{1}{\rho}\omega\,.\end{aligned}$$

This completes the proof in the generality of the assertion. □

As has been the case for the controller strategy from (4.2.11), we can also here show that the values of the regularization estimators (4.2.14) and (4.2.15) converge to zero under the same assumptions as we have stated for the respective transition results for admissibility of the corresponding update steps close to optimal solutions. In particular, this makes fast local convergence as formulated in Theorem 4.1.4 accessible when using this strategy.

Proposition 4.2.6: Limit Behavior of the Remainder Term Strategy

Suppose that all of the assumptions required for the respective transition result from either Proposition 4.1.10 for the original case or Proposition 4.1.13 for the numerically stable formulation hold.

Then, the values of the regularization parameter estimator defined via (4.2.14) or (4.2.15), respectively, converge to zero along the sequence of iterates of the inexact Proximal Newton method.

Proof. As before, we note that the procedure will always provide update steps satisfying the sufficient decrease criterion (4.2.1) and thus global convergence of the ensuing method to an optimal solution is ensured. As we approach the solution, where the additional semi-smoothness assumptions hold, we can thus estimate the values of the corresponding remainder term formulations as follows:

In the generic case of (4.2.14), we remember (4.1.40) stating that

$$f\big(x+\Delta s(\omega)\big) - f(x) - f'(x)\Delta s(\omega) - \frac{1}{2}H_x\big(\Delta s(\omega)\big)^2 = \rho(x,\omega) + o\big(\big\|\Delta s(\omega)\big\|_X^2\big)$$

holds where we additionally denoted $\rho(x,\omega) \coloneqq \frac{1}{2}(H_{x_+(\omega)} - H_x)\big(x_+(\omega) - x_*\big)^2$. This term can be handled by taking advantage of the additional continuity assumption (4.2.13). All in all, we conclude that the absolute value term in (4.2.14) is $o\big(\big\|\Delta s(\omega)\big\|_X^2\big)$ as the current iterate approaches an optimal solution which directly implies the convergence of the estimator to zero.

For the numerically stable formulation from (4.2.15) we can use the corresponding estimates from the proof of the transition result in Proposition 4.1.13. There, we have verified

$$\big\|f'\big(x+\Delta x(\omega)\big) - f'(x) - H_x\big(\Delta x(\omega)\big)\big\|_{X^*} = o\big(\big\|\Delta s(\omega)\big\|_X\big)$$

in the limit of the iterate x to the solution x_* which again provides us with convergence to zero of the estimator. □

We can conclude that the above theoretical framework supports the plausibility stemming from the remainder term illustration for this choice of the regularization parameter and thus rigorously enables its numerical investigation in Section 4.2.3. The decision when to use the

original formulation of the remainder term estimator from (4.2.14) and when the numerically stable alternative defined in (4.2.15) is made under the same deliberations as discussed towards the end of the description of the controller strategy above.

4.2.2 Choice of the Forcing Term

The second algorithmic aspect which we want to take into consideration is the choice of forcing terms for the first inexactness criterion (4.1.13) controlling the relative error of the inexactly computed update steps $\Delta s(\omega)$. As we have pointed out before, local acceleration depends on the forcing terms η tending to zero which is what we will take into account also here.

Generally, there are still various possibilities for the choice of forcing terms with this constraint. In our previous numerical investigations in Section 4.1.6, we have considered a generic null-sequence update of the form

$$[\eta]_1(\eta) \coloneqq 0.6\eta$$

after every successful step computation according to (4.2.1) or (4.2.2). While in theory such a choice eventually yields accelerated convergence, an adaptive alternative is always superior. In particular, the forcing term choice should be able to recognize phases of the algorithm where additional exactness pays off at the expense of computational effort and where we can spare subproblem steps without losing progress in minimization. In particular, this demands the possibility of not only decreasing but also increasing forcing terms throughout the run of the algorithm. For this reason, we will also here take additional information about the objective functional into consideration.

In [24], the authors propose the following two strategies for choosing η_k within the framework of an inexact Newton method to find a root of some differentiable objective function ψ:

$$\eta_k \coloneqq \frac{\left\|\psi(x_k) - \psi(x_{k-1}) - \psi'(x_k)\Delta s_{k-1}\right\|}{\left\|\psi(x_{k-1})\right\|} \tag{4.2.16a}$$

$$\text{or} \quad \eta_k \coloneqq \eta_0 \left(\frac{\left\|\psi(x_k)\right\|}{\left\|\psi(x_{k-1})\right\|} \right)^{\xi} \quad \text{for given} \quad \eta_0 \in]0,1] \ , \ \xi \in]1,2] \, . \tag{4.2.16b}$$

Furthermore, they also add some safeguard values in order to prevent the so-called *oversolving phenomenon*, i.e., the forcing terms from becoming too small too quickly. In [55], this choice has been adapted to a (finite dimensional) Proximal Newton setting via an expression of the form

$$\eta_k \coloneqq \min \left\{ \frac{m}{2} \ , \ \frac{\left\|G_{\hat{f}_{k-1}/M}(x_k) - G_{f/M}(x_k)\right\|}{\left\|G_{f/M}(x_{k-1})\right\|} \right\} \tag{4.2.17}$$

where $M, m > 0$ denote upper and lower bounds on the bilinear forms H_x, G is a generalized composite gradient mapping and $\hat{f}$ is a second order model of the smooth part f. For this implementation, both the knowledge of M and m as well as many evaluations of composite gradient mappings are necessary the problematic nature of which we have already discussed when introducing an existing inexactness criterion at the end of Section 4.1.1. We will now propose two strategies which rather fit the setting that we find ourselves in here.

The Model-Based Approach

The first one is a model-based approach considering the following two aspects: Firstly, the agreement of the model which is used for step computation and the actual non-linearity of the objective functional plays an important role. The better these both coincide, the more exact we want to solve the step computation subproblem since then we can be sure that we will obtain a satisfying update step. In this case, the forcing term η should be chosen very small. If they do not coincide well on the other hand, we want to give some freedom to the inexact step computation and choose η rather large. Strategy (4.2.16a) from above (or (4.2.17), respectively,) greatly regards this aspect.

Secondly, the rate of reduction within the objective functional as in (4.2.16b) should have an impact on our choice in some sense. In [2], the authors propose a new way of choosing η for smooth inexact Newton methods which reflects on both of the aspects discussed above. We adapt their choice to our non-smooth Proximal Newton scenario and show that also here we can take advantage of its beneficial convergence properties.

To this end, we introduce a measure for the actual decrease of F at some x in direction of some update step $\Delta s(\omega)$ relative to the decrease predicted by some arbitrarily regularized second order model $\lambda_{x,\tilde{\omega}}$ via the so-called *reduction quotient*

$$r\colon X \times [0,\infty[\times[0,\infty[\to]-\infty,\infty]\,, \quad r\,(x,\omega,\tilde{\omega}) \coloneqq \frac{F\big(x+\Delta s(\omega)\big)-F(x)}{\lambda_{x,\tilde{\omega}}\big(\Delta s(\omega)\big)}\,. \tag{4.2.18}$$

Thereby, we also provide an indicator for the agreement of $\lambda_{x,\tilde{\omega}}$ and the actual non-linearity of our objective functional F, at least in the direction of the update step $\Delta s(\omega)$.

Note that the parameter $\tilde{\omega}$ in the second order model in the denominator does not necessarily match the one used for the computation of some inexact update step $\Delta s(\omega)$. For coinciding parameters $\tilde{\omega} = \omega$ we find that $r(x,\omega,\omega)$ is just the decrease ratio function $\gamma(x,\omega)$ in the generic case from (4.2.3).

While for $r\,(x,\omega,\tilde{\omega}) \approx 1$ the regularized model and F itself match quite well, this relation deteriorates the smaller the reduction quotient becomes. The case $r(x,\delta x,\tilde{\omega}) \gg 1$ suggests that the models do not coincide but we obtain large descent nevertheless which is tolerable as well. At last, $r\,(x,\omega,\tilde{\omega}) < 0$ would signify objective value increase but should not occur due to the sufficient decrease criterion (4.2.1) for the regularization parameter ω even though the $\tilde{\omega}$ might slightly differ from that one.

Since the above quantities have all been introduced with respect to the original sufficient decrease criterion (4.2.1), we still have to develop a corresponding formulation in the case where we want to achieve additional numerical robustness by taking advantage of (4.2.2) instead. To this end, we define the *alternative reduction quotient* $\tilde{r}\colon X \times [0,\infty[\times[0,\infty[\to \mathbb{R}$ via

$$\tilde{r}\,(x,\omega,\tilde{\omega}) \coloneqq \frac{\big[f'\big(x+\Delta s(\omega)\big)-f'(x)-\big(H_x+\omega\mathfrak{R}\big)\big(\Delta s(\omega)\big)\big]\Delta s(\omega)}{-\tilde{\omega}\big\|\Delta s(\omega)\big\|_X^2}\,. \tag{4.2.19}$$

Apparently, the exact interpretation of this alternative version as a comparison between regularized model and actual non-linearity of the objective function breaks down. On a different note, the respective roles from the corresponding sufficient decrease criterion (4.2.2) justify the definition of both the numerator and the denominator in the above reduction quotient.

The second real-valued argument as before might differ from the damping parameter used for computation of $\Delta s(\omega)$ and determines the "new model" used for subsequent step computation. Due to the diminished significance for the evaluation whether the current step computation model aligns well with the actual non-linearity, we will try to make use of (4.2.19) instead of (4.2.18) only in exceptional cases.

As discussed above, we will now adapt the forcing term η according to the corresponding reduction quotient for intuitively chosen arguments. To this end, we choose parameters $0 < p_1 < p_2 < p_3 < 1$ with $p_1 \in]0, \frac{1}{2}[$ and then determine the trial forcing term $\tilde{\eta}$ according to the following scheme:

At the beginning of step computation for $\Delta s_k(\omega_k)$, the procedure for finding adequate regularization parameters provides us with a trial value $\omega^+ = [\hat{\omega}]_i(\omega_{k-1})$ according to one of the above estimators $i \in \{1, 2, 3\}$ and $\hat{\omega} \in \{\omega, \tilde{\omega}\}$. This determines the second order model λ_{x,ω^+} for the computation of the subsequent step.

Already for the update of the regularization parameter, we have discussed in which scenarios numerically stable formulations should be used and when not. This decision also has to be made here with respect to the contrasting definitions of reduction quotients in (4.2.18) and (4.2.19). We can link this consideration to the one elaborated on for regularization parameters insofar that we now determine a value r^+ depending on whether the original or the alternative sufficient decrease criterion has been used in order to either admit or reject the prior update step, cf. Figure 4.11. According to the respective case for $\hat{\omega}$ before, we now set $r^+ := \hat{r}(x_k, \omega_{k-1}, \omega^+)$ either as in (4.2.18) or as in (4.2.19). With this value at hand, we compute the trial forcing term as $\eta^+ := [\eta]_2(\eta_{k-1})$ according to:

$$[\eta]_2(\eta) := \begin{cases} 1 - 2p_1 & , \ \tilde{r} < p_1 \, , \\ \eta & , \ p_1 \leq \tilde{r} < p_2 \, , \\ \frac{4}{5}\eta & , \ p_2 \leq \tilde{r} < p_3 \, , \\ \frac{1}{2}\eta & , \ \tilde{r} \geq p_3 \, . \end{cases} \tag{4.2.20}$$

Obviously, the shrinking factors $\frac{4}{5}$ and $\frac{1}{2}$ above are arbitrary (< 1) and the result of computational experiments with respect to their effectiveness in [2]. After choosing the forcing term, a trial iterate $\Delta s^+(\omega^+)$ is computed for this η^+. If this trial update does not satisfy the respective sufficient decrease criteria, we adapt ω^+ and recompute the reduction quotient with the most recent information which we have gathered on both the second order model and the actual non-linearity. For this purpose, we use the non-valid trial step $\Delta s^+(\omega^+)$ at the current iterate x_k and compute $r^+ := \hat{r}\big(x_k, \omega^+, [\hat{\omega}]_i(\omega^+)\big)$ again depending on the computational scenario with respect to numerical robustness, i.e., the hat not existing or being a tilde.

This in turn yields a new trial forcing term η^+ determined as in (4.2.20). Once we then obtain a valid update step $\Delta s^+(\omega^+)$, we save $\eta_k = \eta^+$ and $\omega_k = \omega^+$ in order to move on to the next sequence index $k + 1$. The above, rather wordy, description can also be retraced in the illustration of Figure 4.11.

As we have pointed out before, it is crucial for local acceleration that the choice of (η_k) in (4.2.20) yields a null-sequence.

Proposition 4.2.7: Limit Behavior of the Model-Based Forcing Term Approach

Suppose that all of the assumptions required for the respective transition result from either Proposition 4.1.10 for the original case or Proposition 4.1.13 for the numerically stable formulation hold.

Then, the choice for η_k, $k \in \mathbb{N}$, according to (4.2.20) yields a null-sequence for forcing terms.

Proof. It is sufficient to show that the respective decrease quotient $\hat{r}$ converges to anything larger than $p_2 \in]\frac{1}{2}, 1[$ as we approach an optimal solution of problem (3.2.1). This, on the other hand, can be retraced along the lines of the proof of the convergence of the decrease ratio functions from (4.2.3) and (4.2.4).

The only difference is the potentially differing regularization parameter $\tilde{\omega}$ in the denominator model expression. This does not hinder the fundamental arguments of the proof. □

The Regularization-Based Approach

Due to the similar requirements on their convergence behavior towards optimal solutions of the underlying problem, it seems reasonable to tie the choice of η to the one of ω specified beforehand in Section 4.2.1. The previous concept of comparing and assessing the quadratic model is very plausible but in practice might lead to unreasonably small values for η since for very large regularization parameters the models align very well along the very short steps. At the same time, additional exactness in solving the subproblem does not yield the anticipated progress towards the minimizer far away from it.

To this end, we will present another strategy for choosing η which takes into account whether the algorithm is still stuck in the globalization phase where small forcing terms do not lead to acceleration or it has already reached the region of local superlinear convergence where η tending to zero is essential. The behavior of the regularization parameter is a great indicator for the phase of convergence which we find ourselves in.

For this reason, we fix some rather large offset value $\eta_0 \in [0, 1[$ which we then scale according to the current phase of convergence. Then, rather straight-forwardly, we choose the next forcing term for the relative error criterion (4.1.13) as

$$[\eta]_3(\omega_k) \coloneqq \frac{\omega_k}{\omega_{\max}} \eta_0 \tag{4.2.21}$$

where $\omega_{\max}$ denotes the largest regularization parameter which has been employed over the current algorithmic run up until now. This choice implies that as long as the regularization parameter is increasing or constant for subsequent accepted update steps we use the offset value η_0 in order to spare unnecessary subproblem steps during the globalization phase of the algorithm. Once the regularization parameter decreases, indicating the approach of an optimal solution, so does the forcing term within the span $\eta \in [0, \eta_0]$. In particular, as soon as ω might rise again within another globalization phase, the forcing term again rises towards its original value η_0 and the forcing term estimator adaptively recognizes the switch between globalization and anticipated local acceleration phases.

As ω then tends to zero close to the minimizer, η apparently inherits this behavior in the local acceleration phase of the method. We also summarize this property within a proposition:

Proposition 4.2.8: Limit Behavior of the Regularization-Based Forcing Term Approach

The above choice for η_k, $k \in \mathbb{N}$, according to (4.2.21) yields a null-sequence for forcing terms if also the corresponding sequence of regularization parameters tends to zero.

Another aspect, which have decided to add to both adaptive choices of the forcing term, is not of theoretical but rather of algorithmic character when thinking about concrete implementations of our parameter choice: Close to optimal solutions, i.e., once the proximity criterion from (3.2.44) holds, we will always decrease the forcing term at least by a fixed factor once an update step has been rejected by the sufficient decrease criteria. This safeguard-type strategy close to optimal solutions does not disrupt the adaptive behavior of our above choices but rather ensures that admissibility of updates does not deteriorate due to inexactness, and gives a them a rough direction once the globalization phase of our algorithm is over.

4.2.3 Numerical Results

In order to now study the peculiarities in application of the above strategies for choosing algorithmic parameters within our inexact Proximal Newton method, we reconsider the function space problem from the previous investigation of inexactness features in Section 4.1.6.

The Objective Function

In order to shortly recapitulate the setting considered there, we remember that we are looking for some a vector field $u \in H^1_{\Gamma_D}(\Omega, \mathbb{R}^3)$ on the cubical domain $\Omega \coloneqq [0,1]^3$ that minimizes the composite objective functional F defined via

$$F(u) \coloneqq f(u) + \int_\Omega c\|u\|_2 \,\mathrm{d}x$$

for again some parameter $c > 0$ as a weight for the L_2-norm term where the smooth part $f\colon H^1_{\Gamma_D}(\Omega, \mathbb{R}^3) \to \mathbb{R}$ is given by

$$f(u) \coloneqq \int_\Omega \frac{1}{2}\|\nabla u\|_F^2 + \alpha \max\left(\|\nabla u\|_F - 1, 0\right)^2 + \beta \frac{u_1^3 u_2^2 u_3}{1 + u_1^2 + u_2^2 + u_3^2} + \rho \cdot u \,\mathrm{d}x$$

with parameters $\alpha, \beta \in \mathbb{R}$ as well as a force field $\rho\colon \Omega \to \mathbb{R}^3$.

Specifics of the Implementation and Test Scenarios

All in all, the details of the implementation are identical to the ones considered before for this problem except for – obviously – the choice of the regularization parameter and the forcing terms. We use automatic differentiation by adol-C in order to establish the second order model and TNNMG to solve update step computation subproblems. Our main focus is to test whether the theoretical results from above carry over to an actual implementation of the strategies for parameter choice with regard to both algorithmic functionality and computational efficiency when minimizing an actual function space objective with our solver.

As far as the test scenarios within the current section are concerned, we will stay true to the scheme of our earlier investigations in order to provide consistency and sufficient comparability of our results here with respect to findings from previous sections. This means that we will

again choose fixed $\beta = 10$, $c = 10$, and $\rho = -20$ as a load factor for the vector with unit entries. The parameter α, which governs the influence of the non-linear squared max term, will again be varied between 0 and 240 in steps of 40 in order to provide computational scenarios of different difficulty.

The constant determining the subgradient inexactness criterion – where needed – remains unchanged as $\tilde{\omega}_{\max} = 10^{10}$ just like the threshold values $\varepsilon = 10^{-10}$ and $\tilde{\varepsilon} = 10^{-4}$ for the stopping and proximity criteria (4.1.37) and (4.1.46). As before, all results within the current section have been computed after conducting three uniform grid refinements of the cubical domain Ω which results in $8^4 = 4096$ grid elements.

Choices for the Regularization Parameter

We start with the choice of the regularization parameter ω: Before now conducting first actual tests using the different strategies formulated in Section 4.2.1, let us consider quantities and statistics which determine the quality of the respective choice. For the regularization parameter, this deliberation is a quite straight-forward one to make. A good determination strategy chooses ω just as large such that the corresponding sufficient decrease criterion is fulfilled but also just as small such that the second order decrease model is as "pure" as possible. In particular, such a choice implies that we can benefit from unregularized computation as soon as possible in order to make significant progress in terms of minimizing the objective.

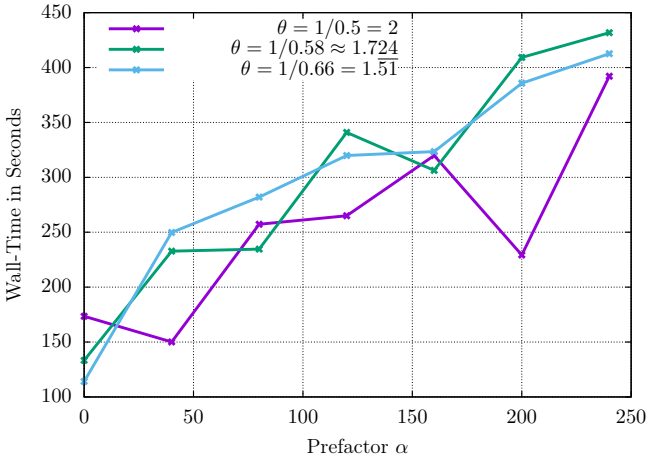
Satisfying the first one of this demands implies that few to no steps have to be rejected due to non-admissibility with respect to the sufficient decrease criterion sparing "outer" iterations in the globalization phase of our algorithm. Fulfilling the second one leads to strongly recognizable local acceleration and a fast approach of the solution in the later stages of the algorithm. In conclusion, we can thus simply summarize that the less wall-time our solver requires to minimize the objective, the better the considered regularization strategy.

In order to keep the focus on the choice of the regularization parameter, we disregard inexact computation of update steps for the corresponding numerical investigations. For both the controller and remainder term approach, we chose the respective safety factor as $\rho = 0.95$, i.e., close to its maximal value of one.

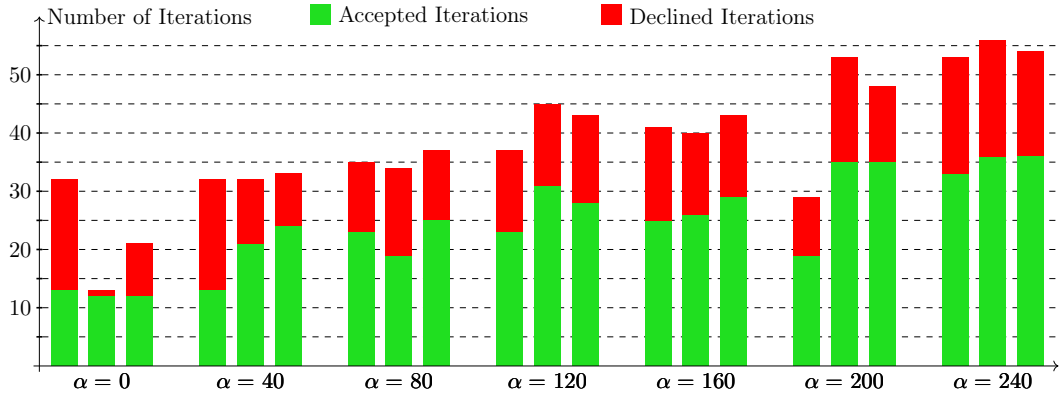
Investigation of the Root Order for the Controller Strategy

However, before comparing the three computational strategies for regularization parameters from Section 4.2.1, let us first consider a rather delicate topic for the controller strategy: the root order within the update formula (4.2.11). Our theoretical deliberations with arguments from control theory for ODEs suggested the choice of $\theta > 3/2$ close to that lower bound. In order to get a grasp on the influence of the choice of this algorithmic parameter, we conduct tests for three different root orders as $\theta_1 = 1/0.5 = 2$, $\theta_2 = 1/0.58 \approx 1.724$, and $\theta_3 = 1/0.66 = 1.\overline{51}$. Since in practice the controller strategy seems to tend to rather drastic increases in the case of non-admissibility of updates, we provide an upper bound in the form that the newly computed regularization parameter should at most be 10 times the previously employed one. This safeguard value does not stand in contrast to the operating principle of the controller strategy but merely provides a boundary for efficient computation. At the same time, we still leave enough space for the philosophy and adaptivity of this idea from ODE step size theory.

As mentioned beforehand, we increase the prefactor $\alpha \in \{0, 40, ..., 240\}$ across one test series. Figure 4.6a shows the wall-times of test runs across this series and Figure 4.6b gives an



(a) Wall-times in seconds.



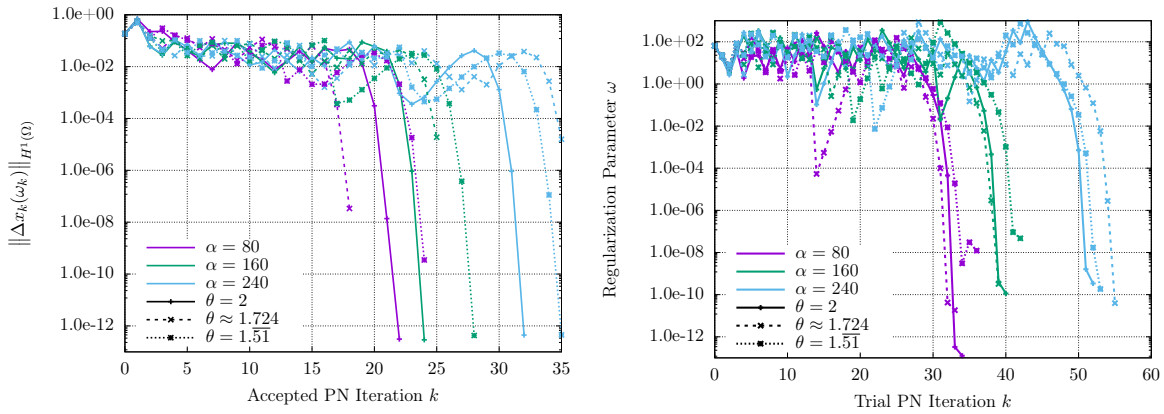
(b) Number of accepted, declined, and total Proximal Newton iterations. For each value of α, the respective numbers are given for three choices of the root order $\theta_1 = 1/0.5 = 2$, $\theta_3 = 1/0.58 \approx 1.724$, and $\theta_2 = 1/0.66 = 1.\overline{51}$ from left to right.

Figure 4.6: Algorithmic Comparison of different root orders across the test series with $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$.

insight on the respective numbers of accepted, declined, and total Proximal Newton iterations required for finding the solution of the minimization problem.

Even though from a theoretical standpoint, θ_3 should provide the best results, our numerical investigations suggest that the largest choice, i.e., $\theta_1 = 2$, is actually the best of the three. This can be retraced to the abovementioned tendency of our adaptive choice to choose rather large regularization parameters in case of failure of the sufficient decrease criterion. While a rather large choice of the root order "flattens the curve" and reduces the trial value for the next ω, small root orders on the other hand rather encourage drastic increases. For local acceleration, however, high root order should work better due to their behavior close to the optimal case $a_{\text{red}} = p_{\text{red}}$ which we want to approach close to optimal solutions. Due to the advantageous performance over the course of the above test series, we have decided to choose $\theta_1 = 2$ for all upcoming tests involving the controller strategy.

From the standpoint of algorithmic functionality, however, we can conclude that the controller strategy, indifferent of the choice of the root order θ, perfectly implements our demands on the regularization parameter choice. Global convergence has been achieved in all test sce-



(a) Correction norms $\|\Delta x_k(\omega_k)\|_{H^1(\Omega)}$.

(b) Regularization parameters ω.

Figure 4.7: Graphs of correction norms and regularization parameters for $c = 10$, $\beta = 10$, $\rho = -20$, $\alpha \in \{80, 160, 240\}$, and different values of the root order θ within the controller strategy. Correction norm plots are not extended for the sake of perspicuity.

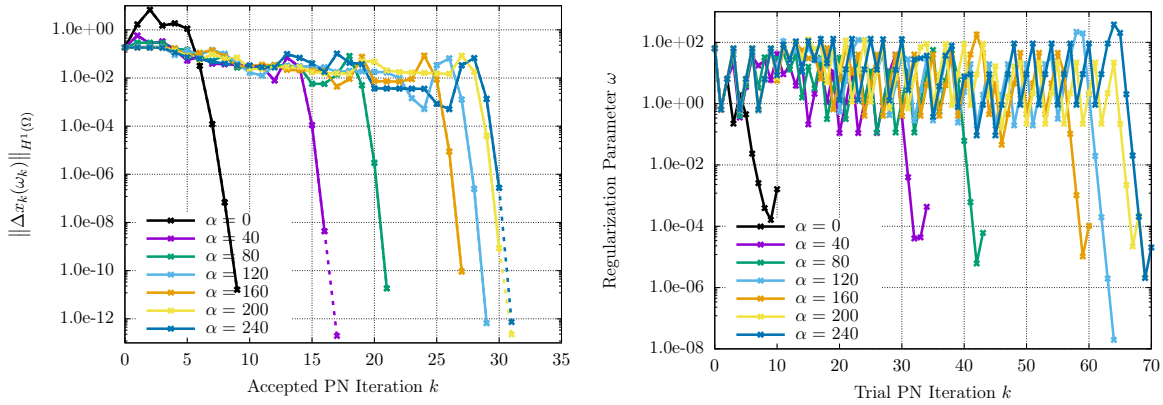
narios and – as can be seen in Figure 4.7 – also local acceleration has been unlocked by the convergence of ω to zero as we approach the optimal solution.

Remainder Term Regularization and Comparison of Strategies

The second adaptive strategy which we have introduced for the choice of the regularization parameter is the remainder term strategy from (4.2.14) and (4.2.15). Since there are no questionable parameters governing the evaluation of this approach, we will first point out the algorithmic functionality of the strategy as displayed in Figure 4.8 and then – without further ado – compare it to the remaining two strategies with regard to computational efficiency. Also for the remainder term procedure we provided a span of values relative to the formerly used regularization parameter in which the new trial ω can be chosen: not more than 10 times but also not less than 0.01 times the previous value.

Figure 4.9a illustrates the wall-times required for solving the minimization problem with (4.1.47) as an objective functional and Figure 4.9b allows us to compare the corresponding numbers of accepted, declined, and total Proximal Newton iterates. From the numerical investigations and resulting data at hand, it is apparent to conclude that the heuristic choice for the regularization parameter yields the most efficient Proximal Newton method – at least across the investigated test scenarios.

Even though one could surely construct a set of model problem parameters and bounds on the adaptive strategies such that these outperform the heuristic choice, this does not stand in conformity with our understanding of scientific rigor. As a consequence, we will simply acknowledge the superior performance of the heuristic strategy and cherish the algorithmic functionality of our adaptive approaches which stands in conformity with our theoretical deliberations from Section 4.2.1. Furthermore, we still believe that adaptive formulations are suited better for generic real-world application problems. Due to its slightly better performance among the two, we will thus choose the controller strategy with $\theta = 2$ as an algorithmic component of our modified method.



(a) Correction norms $\|\Delta x_k(\omega_k)\|_{H^1(\Omega)}$.

(b) Regularization parameters ω.

Figure 4.8: Graphs of correction norms and regularization parameters for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$ using the remainder term strategy for the choice of the regularization parameter. Dashed extensions of the correction norm plots show the last declined step leading to the stop of the algorithm.

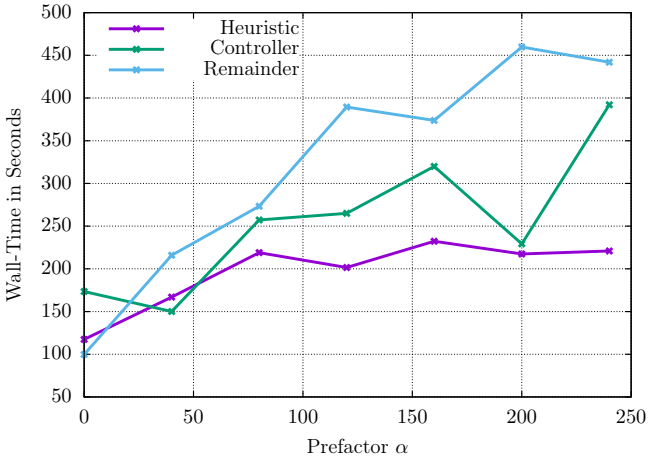
Choices for the Forcing Term

The last algorithmic ingredient which we will investigate before considering the application of our Proximal Newton method to the finite strain plasticity problem is the strategy for choosing the forcing term from the relative error inexactness criterion (4.1.13). In order to keep the focus on the influence of the respective approaches for the control of η, we employ the heuristic regularization strategy across all of the ensuing test runs.

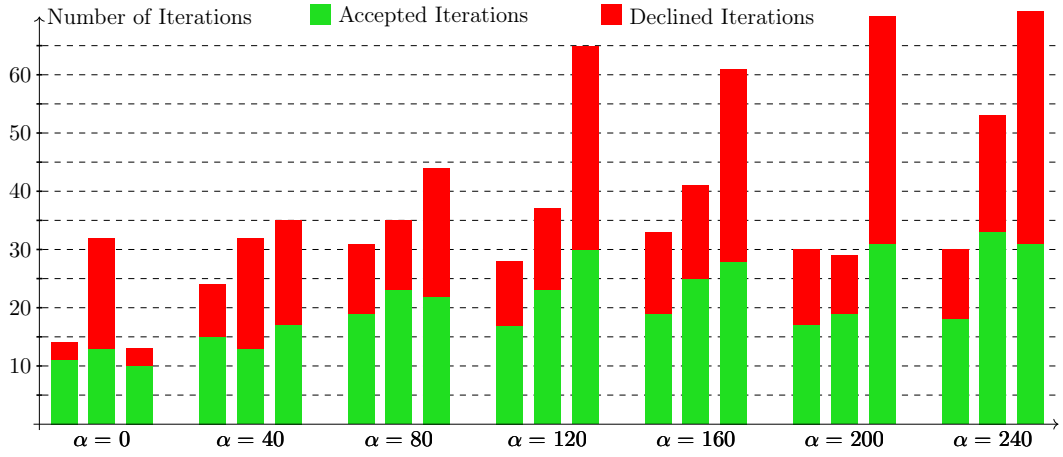
Figure 4.10 shows the graphs of forcing terms across the respective test series which again consists of increasing α from 0 to 240 in steps of 40. The model-based approach from (4.2.20), which is depicted in Figure 4.10a, apparently is able to increase the forcing term after an initial drop but does not exhibit the desired convergence to zero as we approach the solution of the minimization problem. From a theoretical standpoint, this even rules out local superlinear convergence of the ensuing method which can also in practice be retraced in the corresponding correction norm plot in Figure 4.10c.

The regularization-based approach from (4.2.21), however, also in practice exhibits both of our desired properties for the choice of the forcing term. As can be clearly seen in Figure 4.10b, over the course of the runs from our test series, the forcing term both increases during the globalization phase and converges to zero in the local acceleration phase of the algorithm. Even though the forcing strategy simply inherits this desired behavior from the choice of the regularization parameter, this is just what we need from the standpoints of both algorithmic functionality and computational efficiency. As a consequence, we are able to spare unnecessary subproblem solver iterations in the globalization phase of the algorithm while still being able to benefit from local acceleration in the later stages. The local superlinear convergence of the ensuing method is apparent in Figure 4.10d and the gains in computational efficiency can be retraced from the corresponding wall-times compared to the ones of the heuristic approach in Table 4.3.

Let us also note here that the safeguard-like strategy for the forcing term described at the end of Section 4.2.2 only came into play for the model-based approach in the case of $\alpha = 240$



(a) Wall-times in seconds.

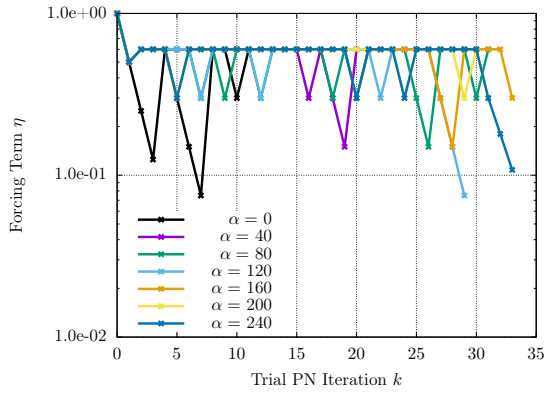


(b) Number of accepted, declined, and total Proximal Newton iterations. For each value of α, the respective numbers are illustrated for the three considered regularization strategies. From left to right: Heuristic, Controller with $\theta = 2$, and Remainder Term.

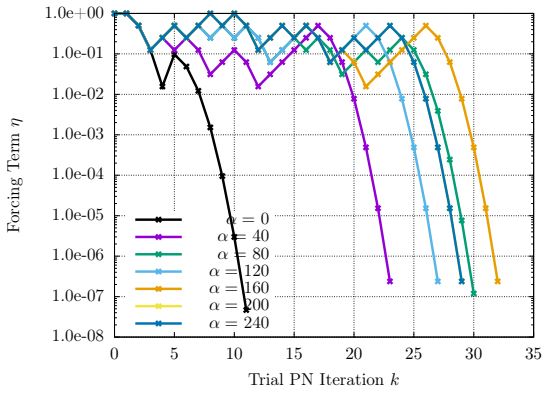
Figure 4.9: Algorithmic Comparison of the three considered regularization strategies across the test series with $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$.

and thus did not affect the numerical investigation of our adaptive parameter strategies here.

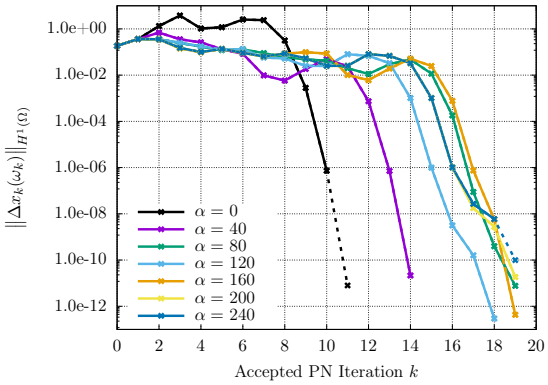
In conclusion, we can say that while the model-based approach does not meet our expectations from theory, the regularization-based approach allows us to boost the computational efficiency of the inexact Proximal Newton method even further. Consequently, we will make the latter strategy a fixed ingredient of our modified method. These findings finalize the numerical investigations of our parameter choice strategies presented over the course of Section 4.2.



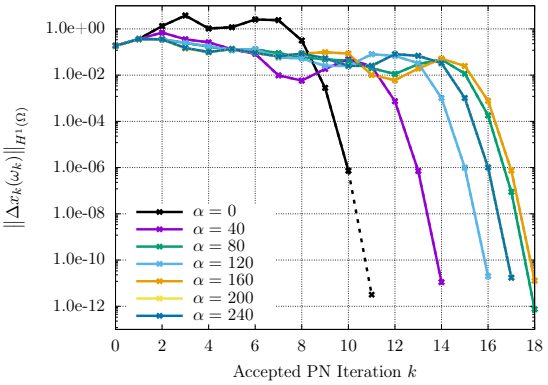
(a) Graphs of forcing terms for the model-based approach.



(b) Graphs of forcing terms for the regularization-based approach.



(c) Graphs of correction norms for the model-based approach.



(d) Graphs of correction norms for the regularization-based approach.

Figure 4.10: Graphs of forcing terms and correction norms across the test series for $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$ using either the model- or regularization-based approach. Dashed extensions of the correction norm plots show the last declined step leading to the stop of the algorithm.

Prefactor α		0	40	80	120	160	200	240
Wall-Times	Heuristic	112.67	152.12	195.11	199.00	201.61	196.66	211.08
	Reg.-Based	111.19	151.24	193.72	174.16	198.21	185.66	188.47

Table 4.3: Wall-times across the test series of $c = 10$, $\beta = 10$, $\rho = -20$ and $\alpha \in \{0, 40, ..., 240\}$ for the inexact Proximal Newton method with either the heuristic or regularization-based approach for choosing the forcing term η in (4.1.13).

4.3 Algorithmic Conclusion

Before we now will expose our algorithm to the computation of solutions for the time-incremental minimization problems in the framework of finite strain plasticity from Section 2.2.4, we will give an overview of it in its final form. This time around, however, we do not resort to the typical algorithmic representation which we have used beforehand but use a more insightful visualization of our computational strategy. Additionally, we recapitulate all inequalities which we require for its formulation afterwards for the convenience of the reader.

The theory behind this algorithmic strategy has been explained elaborately within the preceding sections. The following scheme thus only serves illustrative purposes:

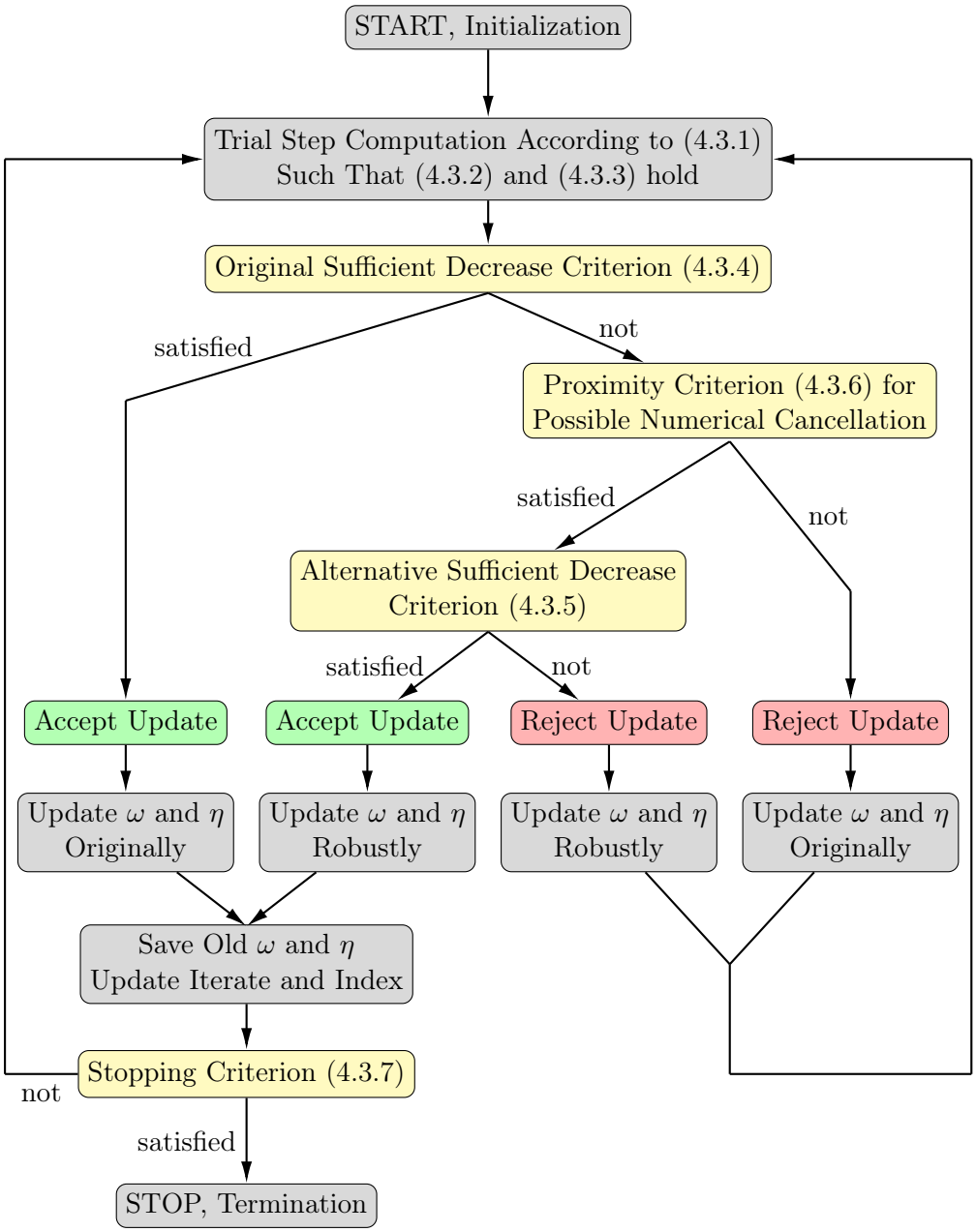


Figure 4.11: Final form of the inexact Proximal Newton method

For the update of the regularization parameter and the forcing term, we have provided formulations both for the original and numerically susceptible setting. In the above algorithmic scheme, we have referred to these respective formulations as updating ω and η "originally" and "robustly". As mentioned beforehand, we now recapitulate all occurring definitions and criteria: Update steps are computed by minimizing a modified second order decrease model

$$\Delta x(\omega) \coloneqq \operatorname*{arg\,min}_{\delta x \in X} \lambda_{x,\omega}(\delta x) \coloneqq f'(x)\delta x + \frac{1}{2} H_x(\delta x)^2 + \frac{\omega}{2}\big\|\delta x\big\|_X^2 + g(x+\delta x) - g(x) \tag{4.3.1}$$

for the regularization parameter $\omega \geq 0$ such that the relative error inexactness criterion

$$\frac{\big\|\Delta x(\omega) - \Delta s^i(\omega)\big\|_X}{\big\|\Delta x(\omega)\big\|_X} \leq \frac{\frac{\theta}{1-\theta}\big\|\delta^i\big\|_X}{\big\|\Delta s^i(\omega)\big\|_X - \frac{\theta}{1-\theta}\big\|\delta^i\big\|_X} \overset{!}{\leq} \eta \tag{4.3.2}$$

holds for some forcing term $\eta \in [0,1[$ as well as subproblem solver increments δ^i, iterates $\Delta s^i(\omega)$, and linear convergence rate $\theta \in]0,1[$. Additionally, the inexact update step has to satisfy the subgradient inexactness criterion

$$\tilde{\omega} \coloneqq -\frac{\big\|f'(x) + \mu\big\|_{X^*}^2}{2\lambda_{x,\omega}\big(\Delta s(\omega)\big)} \overset{!}{<} \tilde{\omega}_{\max} \tag{4.3.3}$$

for an adequately chosen Fréchet-subderivative $\mu \in \partial_F g(x)$ and some large $\tilde{\omega}_{\max} > 0$. The original formulation of the sufficient decrease criterion is given by

$$F\big(x + \Delta s(\omega)\big) - F(x) \leq \gamma \lambda_{x,\omega}\big(\Delta s(\omega)\big) \tag{4.3.4}$$

for the sufficient decrease parameter $\gamma \in]0,1]$ and we introduce the numerically robust formulation via

$$\big[f'\big(x + \Delta s(\omega)\big) - f'(x) - H_x\big(\Delta s(\omega)\big)\big]\Delta s(\omega) \leq \frac{1-\gamma}{2}\omega\big\|\Delta s(\omega)\big\|_X^2 \tag{4.3.5}$$

which uses the same sufficient decrease parameter γ and is checked in case the original decrease criterion fails and additionally the proximity criterion

$$\frac{1+\omega}{1-\eta}\big\|\Delta s(\omega)\big\|_X < \tilde{\varepsilon} \tag{4.3.6}$$

is fulfilled by the current update step for some threshold value $\tilde{\varepsilon} > 0$. Finally, if the stopping criterion

$$\frac{1+\omega}{1-\eta}\big\|\Delta s(\omega)\big\|_X < \varepsilon \tag{4.3.7}$$

is fulfilled for some tighter threshold $\tilde{\varepsilon} > \varepsilon > 0$, the update step computation is terminated and we assume to have reached a solution of the underlying composite minimization problem (3.2.1).

Chapter 5

Application to Finite Strain Plasticity

With all of the theoretical deliberations and numerical investigations with respect to rather simple function space problems from the prior two chapters at hand, it is now time to put our algorithm to its final test: the time-incremental minimization problems from finite strain plasticity which we have already formulated in a suitable fashion in Section 2.2.4.

The Objective Function

There, we have established the structure of the time-incremental minimization problems $(\mathrm{IMP})^{\Pi}$ which is necessary for the application of Proximal Newton methods, i.e., the split of the objective functional

$$F(\boldsymbol{y}, \delta\mathbf{B}) \coloneqq \mathcal{E}\left(t_k, \boldsymbol{y}, \Delta^{+}_{\mathrm{sym}}\big(\mathbf{P}^{k-1}, \exp(\delta\mathbf{B})\big)\right) + T_0 \int_{\Omega} \big\|\delta\mathbf{B}\big\|_F \,\mathrm{d}x \tag{5.0.1}$$

into the smooth stored energy functional and the non-smooth dissipation distance part.

We have already elaborated on possible definitions of the stored energy density and their corresponding shortcomings in view of existence theory in Section 2.2.4. We will go into detail on the respective problem geometries and material models when presenting the scenarios of concrete tests later on. The goal of these tests is to first showcase both the functionality of our method from Chapter 3 in general and the improvements in efficiency by our algorithmic modifications from Chapter 4 for a complicated real-world application problem.

Chapter Outline

In order to achieve this goal, we dissect the chapter in the following way: In Section 5.1, we start out with specifics on the implementation of our solution algorithm for the finite strain plasticity problem which do not depend on the problem geometry and material model at hand. Next, in Section 5.2, we first set up the problem geometry of our "Five" benchmark series and elaborate on the chosen material model, boundary conditions and test scenarios in order to afterwards conduct several algorithmic comparisons with respect to variants of the method presented in this manuscript. At last, in Section 5.3, we will showcase the capabilities of our algorithm on a very demanding and suggestive problem geometry for finite strain plasticity problems: pulling and then releasing one end of a metal paperclip while the other end remains fixed. This will conclude our numerical investigations on the inexact Proximal Newton method.

5.1 Specifics of the Implementation

Let us now lay out the general framework for the concrete implementation of our inexact Proximal Newton method for the finite strain plasticity problem. While many of these aspects can be seen as fairly similar to their counterparts in the rather simple numerical investigations of previous chapters, treating this demanding real-world application also has its peculiarities due to the manifold-structure as already pointed out in Section 2.2.4.

The code which we have used for our numerical investigations has been developed in close collaboration with Patrick Jaap from TU Dresden, who is in particular responsible for most of the plasticity-related implementations like delicate spatial discretization and material model evaluation strategies. A solid foundation for the code has already been established in [92] where he considered the small strain case together with his supervisor Oliver Sander.

However, let us first shortly consider the canonical choices for algorithmic building blocks which we have also used in prior instances of our method: In general, the differentiation of the smooth part of our objective functional is again taken care of by automatic differentiation with adol-C, and solving the discretized update step computation subproblems within the Proximal Newton method is approached via the TNNMG method from Appendix Section A.1.

While the choice of "dynamic" algorithmic parameters ω and η might differ from test to test, their intitial values $\omega_0 = 64$ and $\eta_0 = 0.99$ will remain fixed over the course of all investigations. The same holds true for the general sufficient decrease parameter $\gamma = \frac{1}{2}$. For inexact update step computations, $\tilde{\omega}_{\max} > 0$ governing the subgradient inexactness criterion (4.3.3) is invariably chosen very large as $\tilde{\omega}_{\max} = 10^{10}$. Also $\varepsilon = 10^{-10}$ and $\tilde{\varepsilon} = 10^{-4}$ from the stopping criterion (4.3.7) and the proximity criterion (4.3.6), respectively, will have the same value for all tests.

The initial deformation state is always *undeformed*, i.e., we have $\boldsymbol{y}_0 \equiv \mathrm{id}$ and $\mathbf{P}_0 \equiv \mathrm{I}$ as the corresponding intitial values. We will go into detail on other, rather problem-specific parameters and quantities whenever they start to play an explicit role within our test scenarios.

Spatial Discretization

The aspect which we will elaborate on more deliberately than for our previous numerical investigations is the one of spatial discretization for the main quantities to be computed within the present finite strain plasticity problem. For more details in that regard, we refer to [43], where the discretization techniques described below have been developed and this concern is treated as one of the main subjects of research.

The homotopy step problem, i.e., finding minimizers of (5.0.1) will be discretized in space using different finite element methods. There are three quantities to be discretized: the deformation $\boldsymbol{y}\colon \Omega \to \mathbb{R}^3$, the plastic strain $\mathbf{P}^{k-1}\colon \Omega \to \mathrm{SL}(3)^+_{\mathrm{sym}}$ of the previous load step, and the tangential plastic increment $\delta\mathbf{B}\colon \Omega \to \mathbb{S}^3_0$. Let Ω be discretized by a conforming grid with $n \in \mathbb{N}$ vertices that in particular resolves the Dirichlet boundary Γ_D.

The deformation field $\boldsymbol{y}$ is the easiest to handle. It maps into a vector space and will be discretized using standard first-order Lagrange finite elements. Let $\{\boldsymbol{\phi}_i(x)\}_{i=1,\dots,n}$ be the scalar nodal basis and e_1, e_2, e_3 the canonical basis of $\mathbb{R}^3$. The discrete deformation field is then

$$\boldsymbol{y}_h(x) = \sum_{i=1}^{n} \sum_{j=1}^{3} \bar{y}_{ij} \boldsymbol{\phi}_i(x)\, e_j \tag{5.1.1}$$

with real-valued coefficients $\bar{y}_{ij}$.

For the computation of multiple subsequent homotopy steps, discretizing the plastic strain $\mathbf{P}^{k-1}$ of the previous load step can require more advanced techniques. By the pointwise incompressibility constraint $\mathbf{P}^{k-1}(x) \in \mathrm{SL}(3)^+_{\mathrm{sym}}$ for almost all $x \in \Omega$, the space of all admissible strain fields is non-linear. This problem is trivial to solve if the material model does not contain strain gradient terms ($k_2 = 0$ in (2.2.30)). In that case, $\mathbf{P}^{k-1}$ can be approximated by piecewise constant matrix-valued finite element functions. By the definition of the update operator from (2.2.32), also $\mathbf{P}^{k-1}$ will map to $\mathrm{SL}(3)^+_{\mathrm{sym}}$ for all $k \geq 1$ if $\mathbf{P}^0$ maps to $\mathrm{SL}(3)^+_{\mathrm{sym}}$. For a conforming discretization of the model with a strain gradient term ($k_2 > 0$ in (2.2.30)) we need an approximation space for $\mathbf{P}^{k-1}$ that consists of continuous functions mapping into $\mathrm{SL}(3)^+_{\mathrm{sym}}$. Various such spaces have been constructed by Hardering and Sander [37] under the name of *geometric finite elements*. We use here the so-called *projection-based finite elements* that are defined by taking the space of standard $\mathbb{R}^{3\times3}$-valued Lagrange space and projecting the functions pointwise onto $\mathrm{SL}(3)^+_{\mathrm{sym}}$. For the computation of this projection $\mathbb{R}^{3\times3} \to \mathrm{SL}(3)^+_{\mathrm{sym}}$ and its derivative we use an iterative method described in Appendix A.2.

The third field to be discretized is the field $\delta\mathbf{B}$ of plastic corrections. By definition, $\delta\mathbf{B}$ maps into the tangent space of $\mathrm{SL}(3)^+_{\mathrm{sym}}$ at the identity matrix, or equally, into $\mathbb{S}^3_0$. This is again a linear space, and therefore mappings $\delta\mathbf{B}$ can be approximated by standard first-order Lagrange finite elements. Note, however, that this discretization is not compatible with the one which we have selected for the previous plastic strain $\mathbf{P}^{k-1}$: If $\mathbf{P}^{k-1}$ is approximated by projection-based finite elements, and the strain increment $\delta\mathbf{B}$ by Lagrange finite elements, then the new plastic strain $\mathbf{P}^k$ as produced by the update operator from (2.2.32) will not be a projection-based finite element function.

In principle, this problem could be avoided by replacing the Lagrange finite elements for the approximation of $\delta\mathbf{B}$ by the set of functions that, when used in the update formula from (2.2.32), yield projection-based finite elements. However, these functions are difficult to characterize directly, and we therefore do use Lagrange finite elements for $\delta\mathbf{B}$ and accept the discrepancy as a new source of discretization error. We expect that this error is not larger than the main discretization error.

For the plastic update component $\delta\mathbf{B}$, a basis of the trace-free symmetric matrices is chosen as

$$B_1 \coloneqq \frac{1}{\sqrt{2}}\begin{pmatrix}1&0&0\\0&-1&0\\0&0&0\end{pmatrix}, \quad B_2 \coloneqq \frac{1}{\sqrt{6}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&-2\end{pmatrix},$$

$$B_3 \coloneqq \frac{1}{\sqrt{2}}\begin{pmatrix}0&1&0\\1&0&0\\0&0&0\end{pmatrix}, \quad B_4 \coloneqq \frac{1}{\sqrt{2}}\begin{pmatrix}0&0&1\\0&0&0\\1&0&0\end{pmatrix}, \quad B_5 \coloneqq \frac{1}{\sqrt{2}}\begin{pmatrix}0&0&0\\0&0&1\\0&1&0\end{pmatrix}.$$

This forms an orthonormal basis of $\mathbb{S}^3_0$ under the Frobenius inner product. Therefore, it also defines an isometry between $\mathbb{R}^5$ equipped with the Euclidean norm $\|\cdot\|_2$ and $\mathbb{S}^3_0$ with the Frobenius norm, as

$$\Big\|\sum_{j=1}^{5} a_j B_j\Big\|_F = \|a\|_2$$

holds for all coefficient vectors $a \in \mathbb{R}^5$. For other dimensions than $d = 3$, similar constructions are possible but not relevant for the present treatise. The discrete plastic strain field is

represented by

$$\delta \mathbf{B}_h(x) = \sum_{i=1}^{n} \sum_{j=1}^{5} \bar{b}_{ij} \boldsymbol{\phi}_i(x)\, B_j$$

with scalar coefficients $\bar{b}_{ij}$. Some numerical tricks are included in order to evaluate the objective functional F from (5.0.1). Since order one Lagrangian shape functions are non-negative, the dissipation term can be approximately evaluated as a lumped sum

$$\int_\Omega \big\|\delta \mathbf{B}_h(x)\big\|_F \,\mathrm{d}x \approx \sum_{i=1}^{n} \int_\Omega \big\|\sum_{j=1}^{5} \bar{b}_{ij}\phi_i(x) B_j\big\|_F \,\mathrm{d}x = \sum_{i=1}^{n} \int_\Omega \phi_i(x)\,\mathrm{d}x \big\|\bar{b}_i\big\|_2 =: \sum_{i=1}^{n} \gamma_i \big\|\bar{b}_i\big\|_2 \tag{5.1.2}$$

which enables the usage of efficient solvers like TNNMG, cf. Section A.1, treating these now separated non-smooth parts directly. Due to the symmetry inherent to our spinless approach, the gradient

$$\nabla \mathbf{P}_h(x) = \nabla \left((\mathbf{P}_h^{k-1}(x))^{\frac{1}{2}} \exp(\delta \mathbf{B}_h(x)) (\mathbf{P}_h^{k-1}(x))^{\frac{1}{2}} \right) \tag{5.1.3}$$

is computed explicitly as a sum of three terms using the product rule. The Fréchet-derivative of the matrix square root of a matrix A in direction E can be obtained by setting $s(A) \coloneqq A^{\frac{1}{2}}$ and deriving $s(A)^2 - A = 0$ in direction E. This leads to the Sylvester-like equation

$$s(A)X + Xs(A) = E\,, \tag{5.1.4}$$

which has a unique solution $X \in \mathbb{R}^{3\times 3}$ if all eigenvalues of $s(A)$ are positive [38, Section B.14]. This is in particular the case for $A \in \mathrm{SL}(3)^+_{\mathrm{sym}}$. Then, the square root $A^{\frac{1}{2}}$ is also well-defined. For symmetric A and $d = 3$, the problem (5.1.4) is a linear equation with only 6 unknowns, and hence solved directly. Furthermore, $A^{\frac{1}{2}}$ can be computed by the eigenvalue decomposition of A which is given directly for $d = 3$. The Fréchet-derivative $\nabla \exp(A)[E]$ of the matrix exponential of A in direction E can be easily computed by evaluating

$$\exp \begin{pmatrix} A & E \\ 0 & A \end{pmatrix} = \begin{pmatrix} \exp(A) & \nabla \exp(A)[E] \\ 0 & \exp(A) \end{pmatrix}$$

and extracting the top right block [38, Section 10.6]. The matrix exponential itself is computed by the so-called *Scaling and Squaring Method*, cf. [38, Ch. 10.3]. With these tools at hand, we can properly evaluate the gradient from (5.1.3) by the ensuing product rule expression. For the sake of simplicity, we do not use a projection for the corresponding integral.

In summary, let $\boldsymbol{y}_h$, $\mathbf{B}_h$ and $\mathbf{P}_h^{k-1}$ denote adequate finite element approximations of the deformation field, the plastic update and the previous plastic strain, respectively. The fully discrete increment problem then reads

$$\begin{aligned} &\text{Find } (\boldsymbol{y}_h, \delta \mathbf{B}_h) \text{ that minimizes } \mathcal{E}\Big(\boldsymbol{y}_h, \Delta^+_{\mathrm{sym}}\big(\mathbf{P}_h^{k-1}, \exp(\delta \mathbf{B}_h)\big)\Big) + \mathcal{D}(\exp(\delta \mathbf{B}_h)) \\ &\text{and set } \mathbf{P}_h^k \coloneqq \Delta^+_{\mathrm{sym}}\left(\mathbf{P}_h^{k-1}, \exp(\delta \mathbf{B}_h)\right) \text{ (in a suitable finite element space).} \end{aligned}$$

The minimization problems which are now a part of the above homotopy update formulation are in the following tackled by our (inexact) Proximal Newton method.

5.2 The "Five" Benchmark Tests

The goal for now is to show that single homotopy problems can be solved efficiently by our inexact Proximal Newton method. As a consequence, the first series of benchmark tests for our minimization algorithm has the focus of solving a single increment problem and not chaining them to a whole homotopy of problems which would then govern a time discretization for a time-dependent formulation of a rate-independent system. Later on, we will consider concatenations of different so-called *loading steps and scenarios* which in particular demand the deliberate construction of the spatial discretization as we have described it in Section 5.1.

Section Outline

In order to give a clear overview of how we aim to achieve these goals, we split up the section as follows: At first, in Section 5.2.1, we establish the problem geometry which we will consider as well as describe the grid used to discretize the geometry, clarify the material model together with all related parameters, and specify the boundary conditions which will govern the deformation of the test body throughout our numerical investigations. With this framework at hand, Section 5.2.2 will then consider algorithmic comparisons for the solution process of the single homotopy step problem. To this end, we will contrast the performance of the unmodified version of our Proximal Newton method from Chapter 3 and the refined inexact version with adaptive parameter strategies from Chapter 4.

5.2.1 Problem Geometry and Test Scenarios

We use a test body which has already been studied in the small strain theory in [92]. The body takes the shape of the number 5 which also explains the name of the corresponding benchmark test series referred to as "Five" here. The geometry is coarsely discretized by a grid of 25 cubical grid elements as shown in Figure 5.1. For our numerical test runs, two uniform grid refinements are conducted resulting in 1600 cubical grid elements. The boundary conditions, which govern the deformation of the test body, are placed as follows: The *bottom face* (, where $x_2 = 0$,) is a Dirichlet boundary for the deformation field, i.e., we demand $\boldsymbol{y}(x) = x$ for $x_2 = 0$. External surface loads are applied at the top face $\Gamma_N = [0,4] \times \{7\} \times [0,1]$.[19] For the tensile tests, we will compute a series of different scenarios which are determined by the *loading parameter* $\alpha > 0$ in the energy functional

$$\langle \boldsymbol{\ell}_{\text{pull}}(t), \boldsymbol{y} \rangle \coloneqq \alpha \cdot 10^3 \int_{\Gamma} \boldsymbol{y}_2(x) \, \mathrm{d}S \,. \tag{5.2.1}$$

As a consequence, the load only acts on the second component $\boldsymbol{y}_2$ of the deformation field, i.e., the force applied to the body always points "upwards". Illustratively, the tensile tests conducted here can thus be understood as pulling the steel-like model of the number five upwards with a fixed force until it reaches an equilibrium state. The problem geometry together with boundary conditions and coarse grid discretization are displayed in Figure 5.1.

As we have already mentioned in the above illustrative description, the test body here can be seen as "steel-like". This interpretation stems from our choice to use the St. Venant

[19] Technically, the Neumann boundary Γ_N is determined by the larger set $\partial\Omega \setminus \Gamma_D$ where Γ_D is the bottom face as specified beforehand. Here, however, we refer to Γ_N as the set where non-trivial Neumann forces are applied in our tests.

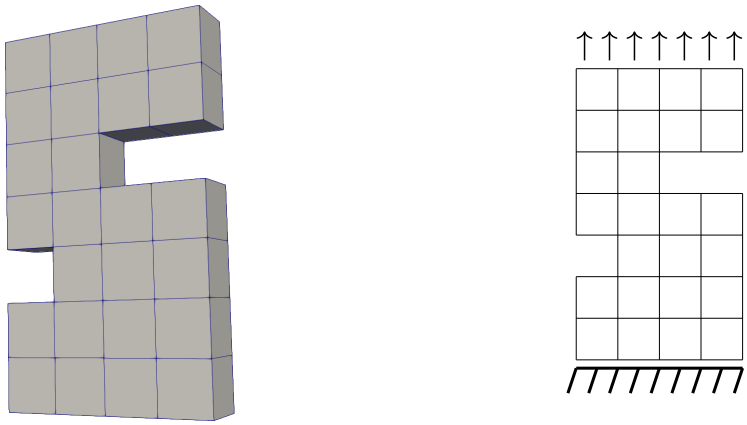


Figure 5.1: Geometry and boundary conditions of the 3D test object "Five".

Kirchhoff material model from (2.2.27) in its simplified form (2.2.29) which we remember as

$$W_{\mathrm{SVK}}(\mathrm{F}) = \frac{1}{2}\lambda|\mathrm{tr}(\mathrm{E})|^2 + \mu\|\mathrm{E}\|_F^2$$

with the Green-Lagrange strain tensor $\mathrm{E} = \frac{1}{2}\left[\mathrm{F}^T\mathrm{F} - \mathrm{I}\right]$. The Lamé parameters for our material model stem from parameter identifications for perfect Prandtl-Reuss elasto-plasticity of a steel alloy specimen conducted in [87, Page 388] which yield a Young's modulus of $E = 206900\,\mathrm{N/mm^2}$ and a Poisson's ratio of $\nu = 0.29$. We can directly translate these parameters to the ones we need for material model determination via

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)} = 8.01937 \cdot 10^4\,\mathrm{N/mm^2} \quad \text{and} \quad \mu = \frac{E}{2(1+\nu)} = 1.107438 \cdot 10^5\,\mathrm{N/mm^2}\,.$$

The yield stress $T_0 = 5 \cdot 10^3\,\mathrm{N/mm^2}$ in the dissipation functional from (5.0.1) is chosen accordingly and the scaling parameters $k_1 = k_2 = 2 \cdot 10^3\,\mathrm{N/mm^2}$ for the kinematic hardening terms from (2.2.30) are of identical value in the same order of magnitude. In particular, we consider the case of linear kinematic hardening as introduced in Section 2.2.4, i.e., we choose $p_{\mathrm{pl}} = p_{\mathrm{gr}} = 2$.

5.2.2 Algorithmic Comparisons

With the problem geometry, material parameters and testing scenarios in mind, we can now turn our attention back to the Proximal Newton algorithms which we have developed over the course of this manuscript.

Comparing "Raw" and "Modified" Proximal Newton Methods

Let us first focus on how we intend to compare the original version of our Proximal Newton method from Chapter 3 with the modified one from Chapter 4 with respect to their algorithmic performance.

In that regard, we refer to the exact Proximal Newton method from Algorithm 10 with the heuristic regularization strategy (4.2.5) as **raw** and to the inexact Proximal Newton method from Algorithm 12 with the controller regularization (4.2.12) ($\theta = 2$) and the

regularization-based forcing (4.2.21) as **modified**.

We will consider the single step homotopy tensile problem from above for $\alpha \in \{1, 2, ..., 7\}$ in (5.2.1) and proceed similar to the numerical investigation of Section 4.1.6 when we added inexactness to our algorithmic framework: At first, we will ensure that both variants obtain the same result. Afterwards, we will compare rather "subjective" indicators of quality for our methods, i.e., graphs of update step norms, regularization parameters, and forcing terms. At last, we will additionally consider "objective" such indicators like the amount of both Proximal Newton and TNNMG update steps as well as wall-time needed in order to find a solution.

For a first illustration of the results achieved by both the raw and modified version of our Proximal Newton method, we refer to Figure 5.2 where the deformed state of the test body at the end of the test is depicted. We only show the solution found by the raw version since the result computed by the modified version relatively differs from that one only by at most $1.31 \cdot 10^{-12}$ in the deformation and not at all in the plastic strain across all three test runs. These relative discrepancies are at every grid point x^k computed by the intuitive formulae

$$\mathrm{err}^P_{\mathrm{rel}}(x^k) := \frac{\left|\left\|\mathrm{P}_{\mathrm{mod}}(x^k) - \mathrm{I}\right\|_F - \left\|\mathrm{P}_{\mathrm{raw}}(x^k) - \mathrm{I}\right\|_F\right|}{\left\|\mathrm{P}_{\mathrm{raw}}(x^k) - \mathrm{I}\right\|_F} \tag{5.2.2}$$

for the relative difference in our plastic strain measure (distance to the identity) and

$$\mathrm{err}^y_{\mathrm{rel}}(x^k) := \frac{\left\|y_{\mathrm{mod}}(x^k) - y_{\mathrm{raw}}(x^k)\right\|_2}{\left\|y_{\mathrm{raw}}(x^k)\right\|_2} \tag{5.2.3}$$

for the displacement where the respective index expressions determine the variant of the Proximal Newton method by which the corresponding grid solution has been computed. The maximal values of the ensuing relative differences for the complete test set are depicted in Table 5.1. Overall, we conclude that both variants of our algorithm obtain the same result within the scenario tested here if both are run until sufficient accuracy is achieved according to the norm stopping criterion (4.3.7).

$\max_{k\in\{1,...,N\}}$	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
$\mathrm{err}^y_{\mathrm{rel}}(x^k)$	$1.81 \cdot 10^{-13}$	$8.51 \cdot 10^{-14}$	$7.56 \cdot 10^{-14}$	$1.39 \cdot 10^{-13}$
$\mathrm{err}^P_{\mathrm{rel}}(x^k)$	0.0	0.0	0.0	0.0

$\max_{k\in\{1,...,N\}}$	$\alpha = 5$	$\alpha = 6$	$\alpha = 7$
$\mathrm{err}^y_{\mathrm{rel}}(x^k)$	$1.12 \cdot 10^{-12}$	$1.31 \cdot 10^{-12}$	$1.45 \cdot 10^{-11}$
$\mathrm{err}^P_{\mathrm{rel}}(x^k)$	0.0	0.0	0.0

Table 5.1: Maximal relative discrepancies of displacement and plastic strain using either the raw or modified variant of the Proximal Newton method across all grid points for the "Five" benchmark series computed via (5.2.2) and (5.2.3).

As a short interpretation of the results depicted in Figure 5.2 from a physical standpoint, one can clearly see that both the deformation of the test body and the plastic strain within

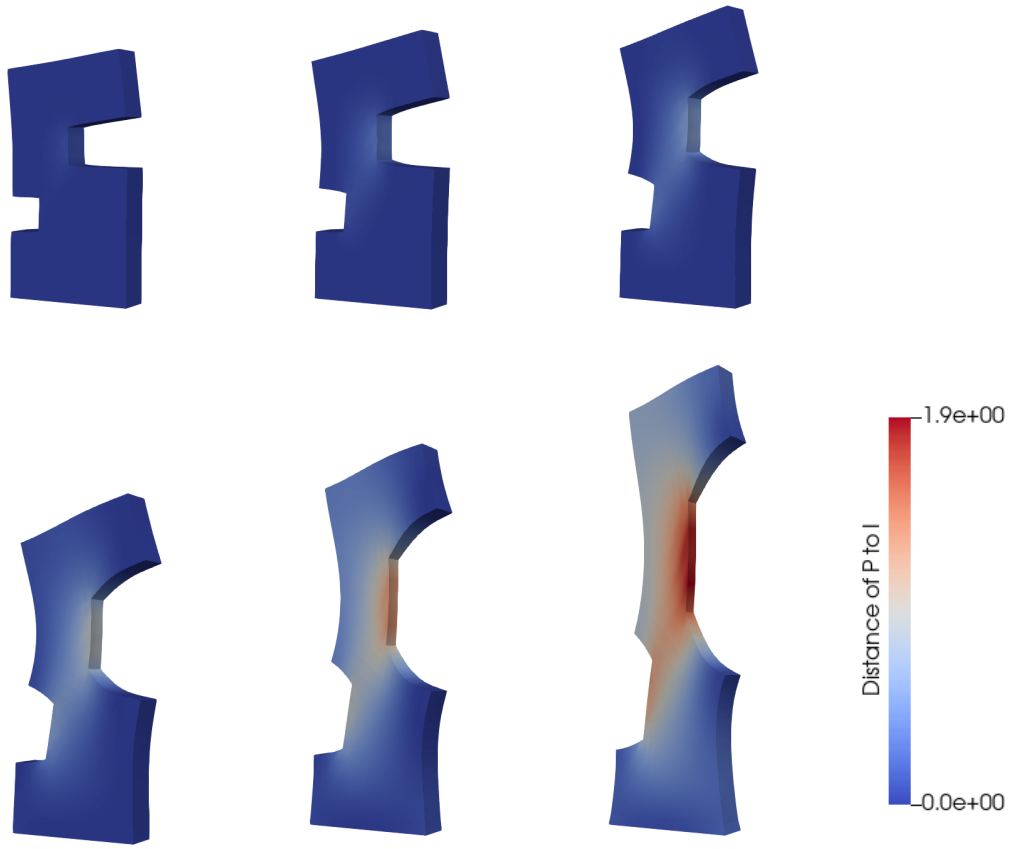
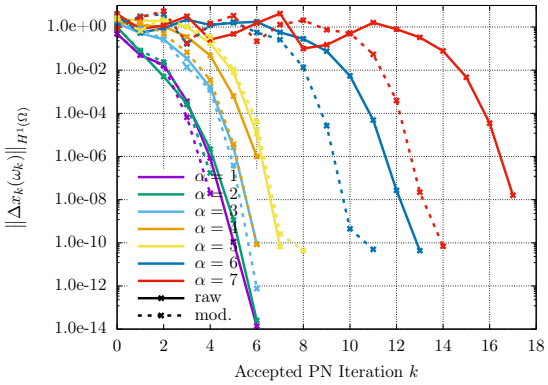


Figure 5.2: Results of the pull test for $\alpha \in \{2, 3, ..., 7\}$ in ascending order.

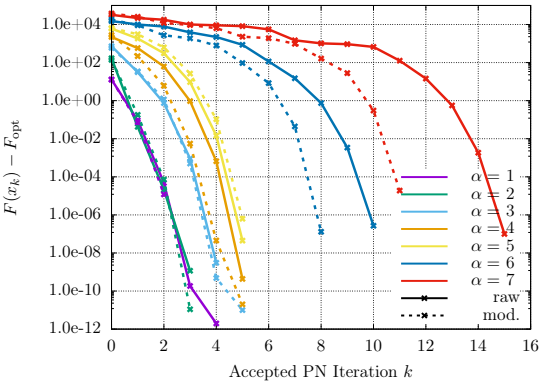
it increase as the force on the Neumann boundary grows larger. For better comparability, the bottom of the test body is fixed to the same height for $\alpha \in \{2, 3, 4\}$ in the top row and for $\alpha \in \{5, 6, 7\}$ in the bottom row, respectively, and the color scales for the plastic strain remain the same across all six pictures. While only a slight deformation and no noticeable plastic strain occur for $\alpha = 1$, these circumstances drastically change along the way to $\alpha = 7$ where all of the test body exhibits non-trivial plastic strain and every part but the Dirichlet boundary on the bottom face is significantly deformed. The volume preserving behavior of our material model becomes apparent insofar that the test body is "truly stretched", i.e., shrinks in width along the deformed parts.

Let us now take a look at the graphs for update step norms, regularization parameters, energy differences and numbers of TNNMG steps per Proximal Newton step across the seven tests conducted here. For higher transparency concerning the algorithmic behavior, we again consider also declined Proximal Newton steps for the regularization parameters and amount of TNNMG steps. The graphs are depicted in Figure 5.3 and can be interpreted as follows: Firstly, both the raw and the modified version clearly exhibit local superlinear convergence both in the correction norms and the energy differences to the optimal value. Secondly, the regularization parameter behaves similarly even though we have to note the increase towards the end of the run for the modifed method and $\alpha \in \{6, 7\}$. There, the rather restrictive nature of the alternative sufficient decrease criterion expresses itself but still helps to obtain global

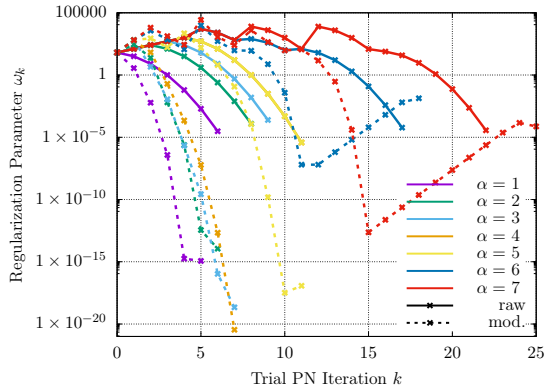
residual convergence as predicted by our theory. We will discuss this phenomenon in detail later. Lastly, the graph for the number of TNNMG steps per trial PN update shows that also within this benchmark series we can spare many (apparently unnecessary) subproblem solver iterations during the globalization phase of our algorithm.



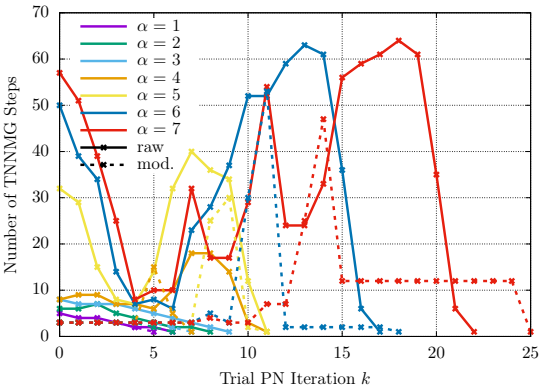
(a) Graphs of correction norms. Plots are not extended for the sake of perspicuity.



(b) Graphs of differences to the optimal energy.



(c) Graphs of regularization parameters.



(d) Graphs of TNNMG iteration numbers.

Figure 5.3: Graphs of correction norms, energy differences, regularization parameters, and TNNMG iteration numbers for the raw and modified method across the "Five" benchmark series.

While the graphs of update step norms, energy differences, regularization parameters, and TNNMG iteration numbers only suggest the advantageous properties of the version of the Proximal Newton method featuring our modifications from Chapter 4, we will now consider some unambiguous measures of quality for the assessment of our algorithmic variants. Table 5.2 gives valuable insight into the statistics behind computations with both the raw and modified version of our Proximal Newton method: It displays the number of accepted ("Acc."), declined ("Decl."), and total Proximal Newton steps required for the respective variant of our method in order to find the solutions displayed in Figure 5.2.

Additionally, both the total number of TNNMG iterations and different shares of wall-time necessary in order to compute the solutions can be found in the table. Here, the "TNNMG" column intuitively refers to the time spent on subproblem solving (including the evaluation of

inexactness criteria), the "Assembler" column lists the time needed to establish the regularized second order models, i.e., mainly to compute gradients and hessians at the respective iterate points. These numbers do not add up to the total wall-time presented in the last column since the latter additionally includes further processes like reading and refining the grid, update procedures within the algorithm, and writing intermediate homotopy outputs and results. In order to give a clear illustration of the corresponding ratios and a comparison between the raw and modified method, the latter information on wall-times is also depicted in Figure 5.4.

α	Variant	PN-Iterations			TNNMG-It.	Wall-Time in sec.		
		Acc.	Decl.	Total		TNNMG	Assembler	Total
1	Raw	7	0	7	21	7.14	219.42	246.15
	Mod.	5	1	6	15	5.14	190.44	212.68
2	Raw	7	2	9	36	19.87	257.29	305.16
	Mod.	5	2	7	18	10.18	218.83	251.32
3	Raw	7	3	10	50	51.53	326.77	415.71
	Mod.	7	1	8	21	21.63	326.11	378.47
4	Raw	7	5	12	110	185.83	471.14	709.70
	Mod.	7	1	8	36	62.15	414.64	514.15
5	Raw	8	4	12	260	640.94	591.49	1295.86
	Mod.	9	3	12	82	215.33	669.20	954.25
6	Raw	14	4	18	576	1662.58	1257.56	3036.69
	Mod.	12	7	19	128	382.75	1116.77	1636.58
7	Raw	18	5	23	773	2427.86	1960.17	4563.79
	Mod.	15	11	26	241	782.66	1573.48	2565.72

Table 5.2: Comparative statistics for the raw and modified variant of our Proximal Newton method with respect to the "Five" benchmark series.

The interpretation of these numbers is straight-forward: While for the rather simple scenarios of $\alpha \in \{1, .., 4\}$ the wall-time required by the assembler dominates the minimization process, for the more demanding problems with $\alpha \in \{5, 6, 7\}$ the algorithmic modifications both with respect to inexact computation of update steps and adaptive parameter choice clearly pay off. Firstly, the gain of efficiency achieved by inexactly computing updates is apparent when taking into account the wall-time spent within the TNNMG subproblem solver. Across the simulations for $\alpha \in \{5, 6, 7\}$, we spared an average of 1116.88 seconds and thereby 70.82% of computational time in that regard. Secondly, also the adaptive choice of parameters has contributed significantly to the shorter total wall-times for our algorithmic runs: On the one hand, the regularization strategy allows for a reduction of the number of accepted Proximal Newton updates in order to find the minimizer. In particular, this implies that the second order model has to be established less often and thereby spares wall-time spent by the assembler. On the other hand, the forcing term strategy adequately controls the accuracy with which the step computation subproblems have to be solved within the respective phases of an algorithmic run. Consequently, we can take advantage of a rather "cheap" globalization phase but still experience local acceleration close to the solution.

Thus, we can conclude that the theoretical deliberations behind our algorithmic modifications in Chapter 4 ultimately payed off and provide us with a minimization algorithm for demanding real world problems which exhibits both algorithmic functionality and efficiency.

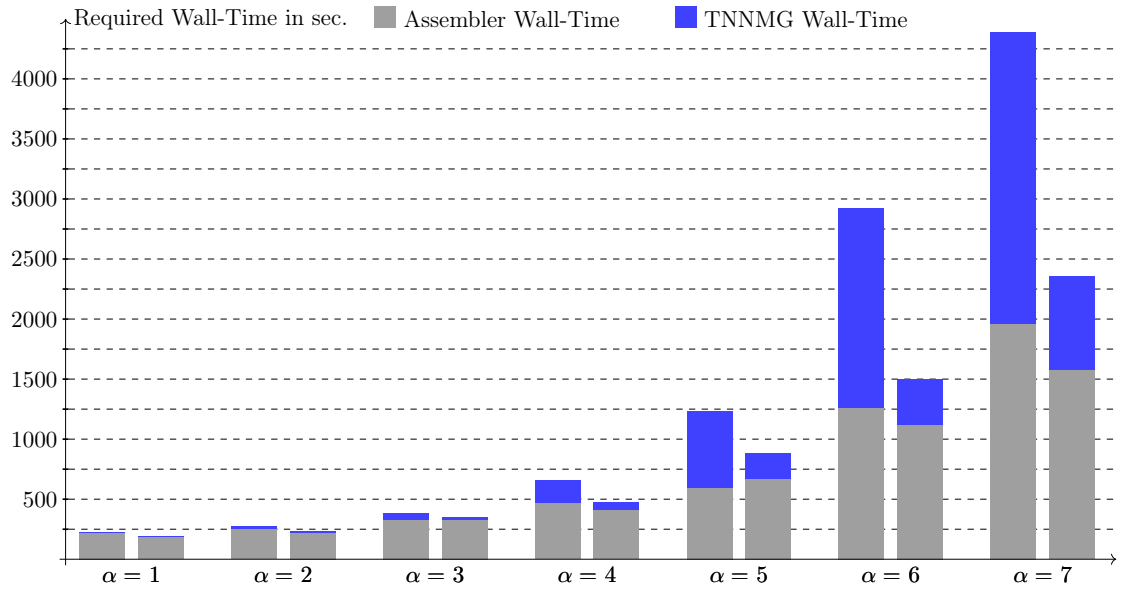


Figure 5.4: Assembler-, TNNMG-, and total wall-times required for algorithmic runs across the "Five" benchmark series. The left bar represents the raw method and the right bar corresponds to the modified variant.

Splitting Into Multiple Homotopy Steps

Another interesting consideration to make is the behavior of our solver with respect to splitting the single step homotopy problem into several "shorter" such steps. To this end, we consider the tensile tests from above for the more demanding scenarios $\alpha \in \{5, 6, 7\}$ but split up the homotopy into five steps in order to compute the same result. Each of the ensuing single homotopy step problems thus should become easier to solve. The idea is to now assess whether splitting up homotopy problems further than the external force intuitively requires is worth from a computational standpoint.

If both discretization schemes yield the same result and thereby significance with respect to the application at hand, the comparison of algorithmic statistics across the respective test runs could provide useful information for the development of simulation strategies for real-world scenarios.

At first, we note that both the plastic strain and the deformation within the test body show non-trivial discrepancies when computed by either one of the ways to discretize the homotopy. We compute the respective relative differences similarly as we have done before for the raw and modified version of our algorithm and this time around take the finer homotopy split as a baseline value. The corresponding formulae are intuitively given by

$$\mathrm{err}^{P}_{\mathrm{rel}}(x^k) := \frac{\left|\left\|\mathrm{P}_{\mathrm{single}}(x^k) - \mathrm{I}\right\|_F - \left\|\mathrm{P}_{\mathrm{split}}(x^k) - \mathrm{I}\right\|_F\right|}{\left\|\mathrm{P}_{\mathrm{split}}(x^k) - \mathrm{I}\right\|_F} \tag{5.2.4}$$

for the relative difference in our plastic strain measure (distance to the identity) and

$$\mathrm{err}^{y}_{\mathrm{rel}}(x^k) := \frac{\left\|y_{\mathrm{single}}(x^k) - y_{\mathrm{split}}(x^k)\right\|_2}{\left\|y_{\mathrm{split}}(x^k)\right\|_2} \tag{5.2.5}$$

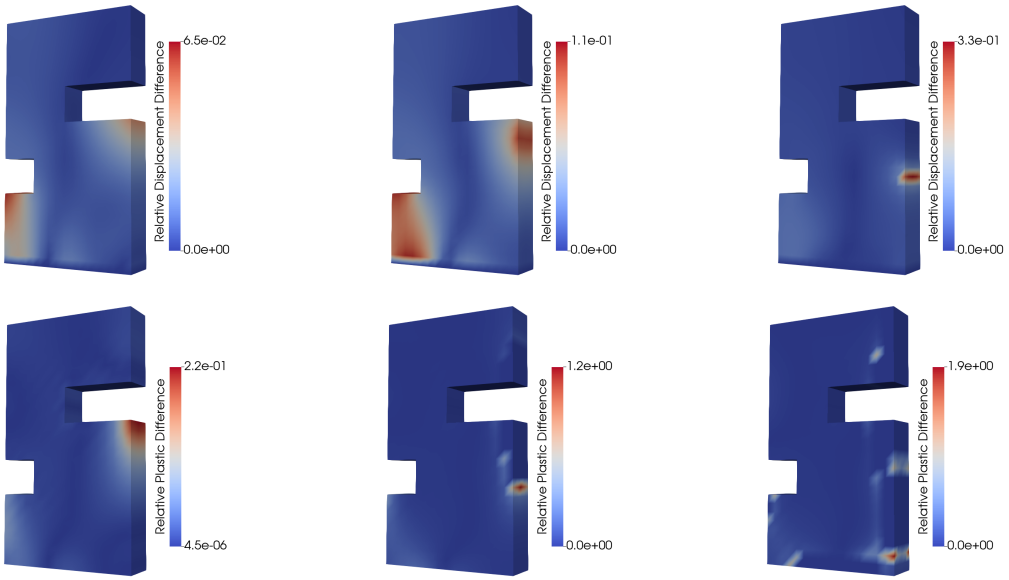


Figure 5.5: Computational relative discrepancies in displacement (top row) and plastic strain (bottom row) according to the formulae (5.2.5) and (5.2.4) for trivial and non-trivial homotopy splitting with (from left to right) $\alpha \in \{5, 6, 7\}$.

for the displacement where again x^k refers to the grid point at which the quantity is computed, and the subscripts "single" and "split" describe the homotopy discretization scheme. The results for these computations are given in Table 5.3 both as the maximal and the average value across the grid. Since these discrepancies between the minimization results are non-trivial, we furthermore illustrate their distribution across the reference configuration in Figure 5.5.

$\max_{k\in\{1,\dots,N\}}$	$\alpha = 5$	$\alpha = 6$	$\alpha = 7$
$\mathrm{err}^y_{\mathrm{rel}}(x^k)$	$3.34 \cdot 10^{-1}$	$1.12 \cdot 10^{-1}$	$7.53 \cdot 10^{-2}$
$\mathrm{err}^P_{\mathrm{rel}}(x^k)$	1.85	1.24	$2.22 \cdot 10^{-1}$

$\mathrm{avg}_{k\in\{1,\dots,N\}}$	$\alpha = 5$	$\alpha = 6$	$\alpha = 7$
$\mathrm{err}^y_{\mathrm{rel}}(x^k)$	$2.01 \cdot 10^{-2}$	$2.11 \cdot 10^{-2}$	$9.62 \cdot 10^{-3}$
$\mathrm{err}^P_{\mathrm{rel}}(x^k)$	$4.34 \cdot 10^{-2}$	$3.12 \cdot 10^{-2}$	$1.33 \cdot 10^{-2}$

Table 5.3: Maximal and average computational relative discrepancies in displacement and plastic strain according to the formulae (5.2.5) and (5.2.4) for trivial and non-trivial homotopy splitting with $\alpha \in \{5, 6, 7\}$.

The non-trivial discrepancies suggest that it is definitely reasonable to use a fine homotopy discretization for significance of results from an application standpoint. In particular, the high maximal relative discrepancies in the plastic strain measure demonstrate the importance of choosing small homotopy step sizes in order to accurately predict the behavior of a test object under Neumann force application. Furthermore, this substantiates the convergence results

from the description of finite strain plasticity problems as rate-independent systems from Chapter 2. The finer we choose the time discretization of the underlying problem, the surer we can be concerning the significance of our computed results. This effect even becomes apparent for the simple, one-directional Neumann forces which we have applied for our computational example here.

Let us now consider the algorithmic comparison of both ways to compute the solution of the homotopy problem: Table 5.4 – just like previous comparisons of that kind – shows the number of accepted, declined, and total Proximal Newton steps as well as TNNMG steps and wall-time shares necessary in order to compute the corresponding result.

α	Variant	PN-Iterations			TNNMG-It.	Wall-Time in sec.		
		Acc.	Decl.	Total		TNNMG	Assembler	Total
5	split	31	13	44	241	446.49	1744.78	2388.22
	single	9	3	12	82	215.33	669.20	954.25
6	split	33	10	43	269	621.65	2107.99	2932.10
	single	12	7	19	128	382.75	1116.77	1636.58
7	split	40	6	46	329	868.71	2920.16	4031.14
	single	15	11	26	241	782.66	1573.48	2565.72

Table 5.4: Comparative statistics for solving the same plasticity problem by either one single or multiple split homotopy steps.

It is easy to see that the discretization scheme of applying the whole one-directional Neumann force to the body within one single homotopy step results in significantly lower required computational time than the split into multiple homotopy steps. The interpretation of this result, however, is not straight-forward: While the coarse discretization in the homotopy is rewarded by quick computation of a result, this result seems to be of lesser quality than the solution computed by splitting the homotopy up into multiple steps. The latter solution has to be interpreted to be more accurate since the theory for rate-independent systems from Chapter 2 predicts convergence of the time-incremental solutions to the energetic one as the fineness of the time discretization tends to zero.

As a consequence, we have the classical trade-off between a fine time discretization, which yields accurate results but requires large computational times, and large homotopy steps which spare wall-time in computation but provide results of inferior physical significance. Investigating the particular interdependence of the time discretization and the computational error is rather a question of solving the time-continuous problem and not of solver development for the increment problems which puts it outside the scope of this manuscript.

On a different note, however, the above recognition lets us come back to the discussion of incorporating plastic spin at the end of Section 2.2.4: Seeing that the choice of a fine time discretization also has significant advantages with respect to solution quality somewhat justifies the assumption of small plastic increments and thereby their symmetry also in the case where plastic spin is incorporated to the problem formulation. This would have been different if the split into multiple homotopy steps above would have yielded the same results but took significantly longer in computation. Then, one would have to split the homotopy only for the sake of choosing symmetric values in the spin formulation. Now, however, a fine time discretization has to be employed in order to obtain significant results regardless of the symmetry of plastic increments which justifies this crucial symmetry assumption a posteriori.

5.3 The "Paperclip" Benchmark Tests

Our second series of benchmark tests for the Proximal Newton algorithm now does not have the focus of comparing methods and displaying algorithmic efficiency of our modifications from Chapter 4. For the last numerical investigations of our methods, we want to display the capabilities of the final modified form of our algorithm in the solution of a finite strain plasticity problem with a more demanding problem geometry and yet another material model. In particular, we will also consider a non-trivial homotopy split of the corresponding (rather simple) time-dependent Neumann forces acting on the test body.

Section Outline

The section itself is structured similarly as the previous one: In the first part, i.e., Section 5.3.1, the problem geometry and its coarse grid discretization are established, the material model together with all related parameters is clarified, and boundary conditions governing the deformation of the test body are specified. Afterwards, in Section 5.3.2, we investigate the numerical solution of the problem via the refined inexact version of our method with adaptive parameter strategies from Chapter 4. As per usual, we to this end consider graphs for significant algorithmic quantities and display the deformed configurations of the test body for the scenario of loading and unloading in two homotopy steps with different pulling forces.

5.3.1 Problem Geometry and Test Scenarios

The test body which we will consider here comes from one of the most natural perceptions of plasticity in everyday life: We will conduct tensile tests of a steel-like paperclip with a total height of 3cm, a total width of 1cm, and a wire diameter of 1mm. The geometry of the model is illustrated in Figure 5.6 and coarsely discretized by 980 prism-like elements. Along

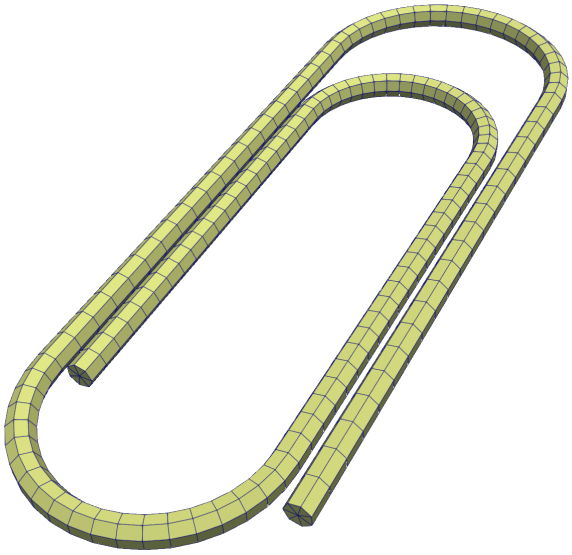


Figure 5.6: Geometry and initial grid of the 3D-paperclip.

the cross sections of the wire there are always 7 triangular elements to model the cylindrical geometry of the wire. For the simulations, one uniform grid refinement is computed resulting

in 7840 grid elements in total. Initially, the paperclip lays in the x-z-plane and zero Dirichlet boundary conditions are placed at the outer face of the wire, i.e., the latter should remain fixed on the "desktop". For the first step within our so-called *binary homotopy step problem*, a scaled Neumann force of the form

$$\langle \boldsymbol{\ell}_{\text{pull}}(t), \boldsymbol{y} \rangle \coloneqq \alpha \cdot 25 \int_{\Gamma} \boldsymbol{y}_2(x) \, \mathrm{d}S \,. \tag{5.3.1}$$

for the loading parameter $\alpha > 0$ is applied to inner end of the wire which thus constitutes the Neumann boundary Γ_N. The load only acts on the third component y_2 of the deformation field which represents pulling the inner end of the wire upwards, i.e., perpendicular to the desktop in our illustrative example. The *loading step* as before ends when the paperclip reaches an equilibrium state induced by the applied Neumann force. Afterwards, in the so-called *unloading step* we completely let go of the force and investigate to which extent both the elastic and plastic deformation of the paperclip remain.

As we have already mentioned beforehand, we still consider a "steel-like" material but this time around will use a different stored energy functional for modeling the material properties. To this end, we will use an approximation technique from [14, Theorem 4.10-2] which allows us to give a representation of our previously used St. Venant Kirchhoff material model in the form of a polyconvex Mooney-Rivlin formula as introduced in (2.2.25). More precisely, we use

$$W_{\text{MR}}(\mathrm{F}) \coloneqq a\big\|\mathrm{F}\big\|_F^2 + b\big\|\mathrm{cof}(\mathrm{F})\big\|_F^2 + c\det(\mathrm{F})^2 - d\ln(\det(\mathrm{F})).$$

In order to favor volume-preserving behavior, we set $d = 2a + b + c$, such that for hydrostatic deformations $F = \rho\mathrm{I}$ the energy $W_{\text{MR}}(\rho\mathrm{I})$ is minimal for $\rho = 1$. By fitting the remaining parameters $a, b, c > 0$ to the Lamé coefficients λ and μ from above, we can thus reach an approximation of the corresponding material model up until cubical terms in the Frobenius norm of the Green-Lagrange strain tensor E. Thus, by choosing $c = 10^4\,\mathrm{N/mm^2}$ and from there computing $a = 4.5323475 \cdot 10^4\,\mathrm{N/mm^2}$ as well as $b = 1.0048425 \cdot 10^4\,\mathrm{N/mm^2}$, we obtain

$$W_{\text{MR}}(\mathrm{F}) = W_{\text{SVK}}(\mathrm{F}) + O\big(\big\|\mathrm{E}\big\|_F^3\big) \quad \text{for } \mathrm{E} \coloneqq \frac{1}{2}\big(\mathrm{F}^T\mathrm{F} - \mathrm{I}\big)\,.$$

The yield stress within the dissipation functional from (5.0.1) is now set to $T_0 = 2 \cdot 10^3\,\mathrm{N/mm^2}$ and the scaling parameters $k_1 = k_2 = 2 \cdot 10^3$ for the kinematic hardening terms from (2.2.30) are of identical value in the same order of magnitude. In particular, we as before consider the case of linear kinematic hardening as introduced in Section 2.2.4, i.e., we choose $p_{\text{pl}} = p_{\text{gr}} = 2$.

5.3.2 Numerical Results

Having established the problem geometry, material model, and general testing scenario over the course of our above elaborations, we will now consider the solution of the ensuing binary homotopy step problem for loading with fixed parameters $\alpha > 0$ and subsequent unloading in order to simulate the behavior of the steel paperclip. Quite intuitively, we will refer to the test scenarios described above for $\alpha \in \{1, ..., 4\}$ as the "Paperclip" benchmark series.

For the interpretation of the corresponding results, let us start with Figure 5.7 which depicts the deformation of the paperclip for the pulling scenarios $\alpha \in \{2, 4\}$ both in the loading and the unloading step. It is apparent that both the displacement and the plastic strain increase as the Neumann force is doubled. In both scenarios, we recognize that after

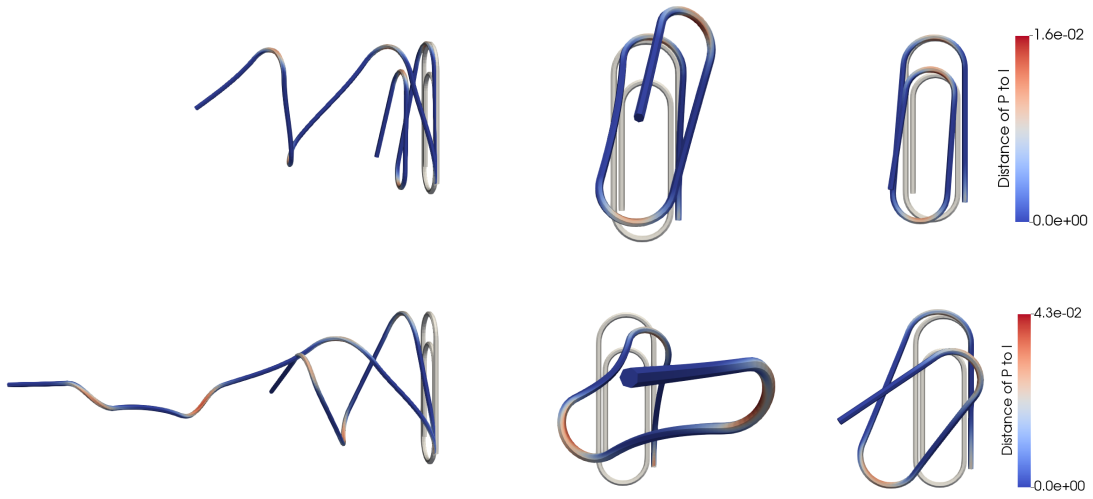


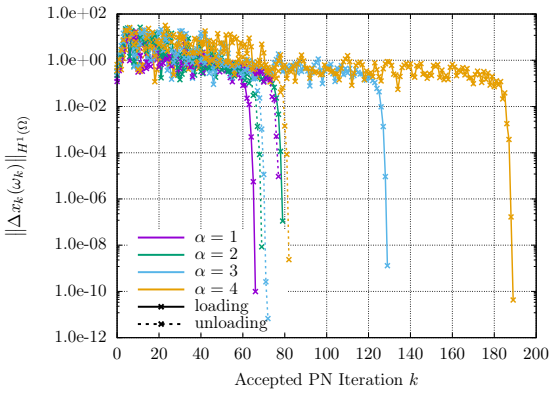
Figure 5.7: Results of the paperclip deformation simulation for $\alpha = 2$ (top row) and for $\alpha = 4$ (bottom row). From left to right: side-view for both loading and unloading, front-view for loading, front-view for unloading. The reference configurations are displayed in solid color.

unloading the two variables determining the deformation of the test body have diminished but still remain significant. This demonstrates the irreversible nature of plastic deformation and thereby stands in conformity with the deliberations made for material modeling in Section 2.2. Thinking of a "real" paperclip, however, one might expect the displacement to diminish less while unloading, i.e., letting go of the previously employed force to the inner end of the paperclip. Since we are rather interested in the behavior of our solution algorithm than in the one of the material, we put this "problem" of choosing adequate material parameters out of the scope of our elaborations here.

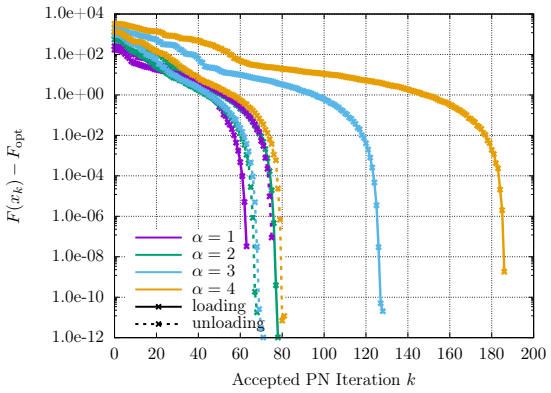
As far as the behavior of the solver is concerned, we take a look at Figure 5.8 which contains all of the relevant information in order to evaluate the algorithmic runs for this benchmark series. The superlinear convergence of the method is apparent when considering the graphs for correction norms and energy difference to the optimal value both in the loading and the unloading step. In particular, the globalization phases are significantly longer for this demanding problem than for previously considered scenarios. Once the region of local convergence is encountered by the algorithm, however, the acceleration allows us to quickly approach the optimal solution of the problem.

The investigation of graphs for regularization parameters and forcing terms from the remaining two images in Figure 5.8, on the other hand, reveals the rather restrictive nature of the alternative sufficient decrease criterion. We have already gotten a glimpse of that phenomenon for the more demanding scenarios of the "Five" benchmark series. In particular, we perceive the increase of regularization parameters towards the end of the algorithmic run in four of the eight homotopy steps considered above – in the loading step for $\alpha \in \{2, 3, 4\}$ and in the unloading step for $\alpha = 1$.

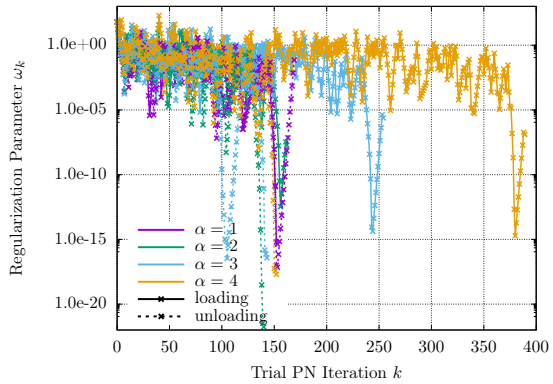
As we take a closer look at – for example – the loading step for $\alpha = 4$, we can see that the regularization parameter again starts decreasing just before update step computation ends due to the stopping criterion. We interpret this behavior of ω in the way that – in particular for very demanding scenarios – the neighborhoods of optimal solutions from the respective



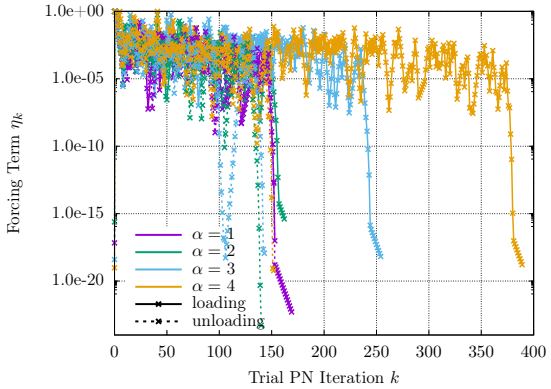
(a) Graphs of correction norms. Plots are not extended for the sake of perspicuity.



(b) Graphs of differences to the optimal energy.



(c) Graphs of regularization parameters.



(d) Graphs of forcing terms.

Figure 5.8: Graphs of correction norms, energy differences, regularization parameters, and TNNMG iteration numbers for the raw and modified method across the "Paperclip" benchmark series.

transition results for either the original or the alternative sufficient decrease criterion in Propositions 4.1.10 and 4.1.13 significantly differ in size. While this region of general admissibility for undamped update steps has clearly already been entered in the aforementioned scenario around trial step $k = 370$ for the original formulation, this is not the case for the numerically robust version.

As a consequence, the regularization parameter is decreased further and further by the original procedure – up until the point where the corresponding sufficient decrease criterion can not be evaluated any more due to numerical cancellation. Then, as intended, we switch to the robust formulation but find ourselves outside of the neighborhood of the solution which allows for unregularized computation by this sufficient decrease criterion. For that reason, we have to increase the regularization parameter again in order to continue global convergence as described in Section 3.2.6.

The cause for these symptoms, however, is not an inherently wrong formulation or defective algorithmic development but rather the insufficient quantifiability of regions for undamped update steps in computationally demanding scenarios. Due to the general framework of our

formulations with respect to differentiability and convexity, this is a drawback which we have to acknowledge. As we conclude the algorithmic behavior of our method, however, we can definitely say that it passed the test also in the most demanding scenario at least considered over the course of the present treatise.

α	Phase	PN-Iterations			TNNMG-It.	Wall-Time in sec.		
		Acc.	Decl.	Total		TNNMG	Assembler	Total
1	Loading	67	87	154	453	314.4	16292.3	20659.44
	Unloading	78	92	170	344	232.0	19025.2	24603.3
2	Loading	80	83	163	630	1240.6	28177.4	36318.7
	Unloading	70	71	141	288	197.2	16989.8	20420.6
3	Loading	130	125	255	1221	3640.4	55379.9	68187.8
	Unloading	73	71	144	553	410.9	18461.0	24533.1
4	Loading	190	200	390	2254	8004.9	92608.9	115347.5
	Unloading	83	70	153	595	542.4	24459.7	32843.6

Table 5.5: Algorithmic statistics for the solution of the binary homotopy step problems from the "Paperclip" benchmark series.

As a last deliberation, we consider the statistics for the "Paperclip" benchmark series which are listed in Table 5.5. Two central perceptions strike the eye as we take a look at the corresponding numbers of iterations and wall-times across all values for the Neumann force prefactor: Firstly, the complexity of the problem becomes apparent by the order of magnitude in which the quantities are located – in particular in comparison with the scenarios considered in previous numerical investigations of our method. Secondly, we recognize that the time required for assembling subproblems again constitutes the determining share of wall-time in all computations. On the one hand, this shows that considerable room for improvement is still present in that regard which lays beyond the scope of the present treatise. On the other hand, this allows for the conclusion that the remaining part of computations necessary for finding a solution is handled quite well by the procedure which we have intended to optimize over the course of the manuscript.

Chapter 6

Conclusion and Outlook

Let us now shortly reflect on both the achievements within the present treatise and possible enhancements of our theory developed here in order to either improve the established results or cover corresponding topics in greater generality:

We have developed a Proximal Newton method for function space problems which works under very general assumptions with respect to convexity and differentiability of the objective functional. Global convergence of the method does not rely on the strong convexity of either the smooth or the non-smooth part of the objective due to a quadratic norm regularization strategy within the update step computation subproblems. Furthermore, we have been able to verify local accelerated convergence of the method in case the functional to be minimized exhibits additional convexity properties close to stationary points of the problem. Establishing this result has been made possible by the consideration of adequate semi-smoothness assumptions together with a generalized interpretation of scaled proximal mappings as operators originating in the dual space of the underlying Hilbert space. In particular, our theory depends on the choice of the second order bilinear forms H_x as Newton-derivatives of the Lipschitz-continuous derivative of the smooth part. This rather restrictive property can under adequate assumptions be generalized to the so-called *Dennis-Moré* condition (3.1.21) which has not been pursued in our context.

For the transition to local convergence of our method, we have introduced the novel concept of second order semi-smoothness for continuously differentiable operators. Similar to existing notions of semi-smoothness, the definition uses an approximation property which is adequately lifted to the second order level for the evaluation of corresponding models in the context. Furthermore, we have established a general calculus for the concept and have given sufficient requirements for the second order semi-smoothness of adequately defined superposition operators. These theoretical aspects of the manuscript can be augmented by considering the question of the concrete choice of the generalized second order differential for a given example and providing strategies to then show second order semi-smoothness with respect to it. The treatment of these aspects has been pursued for "first order" semi-smoothness in the general example of non-linear complementarity problems in [111] and following a similar approach looks promising also in the present scenario.

Coming back to algorithmic deliberations, our alternative sufficient decrease criterion for numerical robustness allows for very accurate identification of limit points due to computational evaluability also for very small update steps. This modification of the basic Proximal Newton algorithm can be considered a powerful tool which significantly improves the convergence behavior for concrete implementations of our method. In contrast to these advantageous

properties and the well-established theory for the alternative formulation, concrete implementations sometimes encounter problems with satisfying the criterion as we have already elaborated on towards the end of Section 5.3. To this end, a more extensive investigation regarding the switch of corresponding characterizations for sufficient decrease in demanding scenarios might provide a better understanding and even yield algorithmic improvements of the concept.

Furthermore, we have modified our method for algorithmic efficiency with the introduction of inexactness to update step computations. For that reason, we have payed particular attention to peculiarities with respect to both the composite structure of the minimization problem and the function space setting which we find ourselves in. In particular, we have designed inexactness criteria which can be evaluated efficiently, save a considerable amount of computational effort, and preserve the favorable convergence properties of the exact method.

As far as the choice of algorithmic parameters is concerned, we have introduced adaptive strategies which use the structure of the underlying minimization problem. The controller strategy takes inspiration from time step size choice for numerics of ODEs and works with estimated relative error quantities between the second order model and actual non-linearities of the composite objective functional in order to adjust the regularization parameter for the next trial step computation. Similarly, the remainder term strategy in an intuitive way estimates the regularization parameter as a suitable prefactor for capturing non-linearities which are not incorporated into the second order model used for update step computation. Both of these approaches conceptually work very well in practice but are slightly outperformed by the heuristic alternative in the test scenarios which we have considered. This behavior should be investigated across a greater variety of applications which might yield a better understanding of possible drawbacks and enable the determination of optimal meta-parameter configurations.

The main application in mind for the newly developed function space algorithm has been the solution of time-incremental minimization problems for rate-independent finite strain plasticity. We have explained all of the underlying concepts for this specific problem in sufficient generality and have elucidated the contribution of the individual components of the model to the existence of solutions to time-incremental and time-continuous problems. Furthermore, we have elaborated on their importance for the significance of the ensuing physical description for real world phenomena. As is to be expected, there is a conflict of interest between theory and application but we have tried to pursue both aspects equally and bring them into accordance for our formulation. An augmentation of our work could rigorously incorporate the concept of plastic spin with respect to both modeling theory of elasto-plastic media and the implementation of the corresponding solution algorithm.

With the restriction to spinless plasticity, however, we have successfully reformulated the time-incremental minimization problems of the corresponding rate-independent framework into an optimization problem which can be handled by our Proximal Newton method. The results from the application of the latter to that problem have fulfilled our expectations and verify our procedure as an efficient way to simulate the behavior of plastic materials at finite strain. Room for improvement is still present in assembling second order models of the objective function in short time which has not been considered in this manuscript.

In order to make the scheme of our Proximal Newton method applicable to further problem classes, the incorporation of equality constraints to the composite minimization problem suggests itself to be a natural extension. Even though this augmentation will surely be non-trivial both with respect to theoretical deliberations and numerical implementations, composite step methods have proven themselves in that regard across a wide field of applications, cf. e.g. [58, 94], and constitute a promising perspective for future scientific work in that direction.

Appendix A

Specifications of the Implementation

The Appendix chapter consists of the following contents: At first, in Section A.1, we give an elaborate description of the TNNMG method which is used in order to solve the update step computation subproblems across all numerical investigations of the manuscript. Afterwards, Section A.2 concerns the projection algorithm used for mapping matrices to elements from $\mathrm{SL}(3)^+_{\mathrm{sym}}$ in the spatial discretization for the finite strain plasticity problem in Section 5.1. The specifications on the test machine, on which the calculations have been conducted, can be found in Section A.3. At last, we give instructions for recomputing the data evaluated across the present treatise in the data availability statement of Section A.4.

A.1 A Truncated Non-Smooth Newton Multigrid Method

As we have pointed out several times throughout the derivation and numerical investigation of our inexact Proximal Newton method, a central ingredient for concrete implementations of the algorithm is a (at least linearly convergent) subproblem solver for finding update steps. For this algorithmic component, we want to use as much structure of our composite objective functional as possible. For the applications of our method which we have mind, the main structural peculiarity to be exploited is so-called *block-separability*.

The subproblem solver which we take advantage of is the so-called *Truncated Non-smooth Newton MultiGrid method* (TNNMG) which constitutes a robust and efficient solution algorithm for a wide range of block-separable convex minimization problems. Originally, it has been designed with the intent to use it for such problems stemming from discretizations of non-linear and non-smooth partial differential equations. We will shortly elaborate on its conceptual idea and main algorithmic components here, and refer to the overview from [33] for a detailed description and convergence analysis. Our deliberations here in particular mirror the ones from [33].

The kind of minimization problems which TNNMG is tailored to is of the following form: Given an objective functional $\mathcal{J}\colon \mathbb{R}^n \to]-\infty, \infty]$, we assume the additive split

$$\mathcal{J} = \mathcal{J}_0 + \varphi \tag{A.1.1}$$

where $J_0\colon \mathbb{R}^n \to \mathbb{R}$ is coercive and (at least) continuously differentiable, and $\varphi\colon \mathbb{R} \to]-\infty, \infty]$ is block separable. Block separability is characterized by the existence of functionals $\varphi_i\colon \mathbb{R}^{n_i} \to]-\infty, \infty]$ for $i \in \{1, ..., M\}$, $M \in \mathbb{N}$, that are convex, proper, lower semi-continuous, and

continuous on their domains such that

$$\varphi(v) = \sum_{i=1}^{M} \varphi_i(v_i) \tag{A.1.2}$$

holds for all $v \in \mathbb{R}^n$ with correspondingly restricted $v_i \in \mathbb{R}^{n_i}$. Apparently, also $\sum_{i=1}^{M} n_i = n$ is necessary here. For a detailed analysis as in [33], it is necessary to give precise definitions of *compatible* Euclidean space decompositions and *block separability with respect to* such decompositions but for our introductory elaborations here the above formulation is sufficient.

While in previous investigations, TNNMG has been successfully employed for the solution of contact and friction models in solid mechanics (cf. [50],[81]), certain models of porous media flow (cf. [8]), Allen-Cahn-type phase-field models (cf. [48]), and even small strain elasto-plasticity (cf. [92]), we cannot directly solve a discretized version of our global homotopy step problem from Section 2.2.4 using TNNMG since, in general, convexity of the stored energy functional $\mathcal{E}$ is not ensured. Therefore, another method is necessary to create a series of regularized convex subproblems which can be handled by TNNMG even though they do not have the motivational PDE-background. Additionally, TNNMG is a purely algebraic solver formulated for finite dimensional minimization problems only. To this end, we consider the thoroughly described inexact Proximal Newton method as an iterative regularizing procedure to fulfill this task.

The TNNMG method can be viewed as a combination of local relaxation methods with a generalized Newton approach. Generally, a non-linear pre-smoother is followed by an inexact linear corrector step of flexible definition – even though typically given by one multigrid iteration. In particular, it achieves multigrid-like convergence behavior (, i.e., mesh-independent convergence rates and linear time complexity,) on a wide range of difficult non-linear problems without regularizing them or involving parameters that would need to be selected manually.

Let us now give an algorithmic overview of the TNNMG method as introduced in [33, Section 3] and give a short explanation of each algorithmic component afterwards. There, for given initial iterate $u^0 \in \mathrm{dom}\mathcal{J}$ and iteration number $\nu \in \mathbb{N} \cup \{0\}$, one update step is determined as follows:

As can be retraced in the scheme from Algorithm 13, the update step computation for TNNMG generally consists of three main components: The first one is *non-linear pre-smoothing* which can in this formulation be understood as a *non-linear Gauß-Seidel iteration*. More specifically, in (A.1.3) the objective functional $\mathcal{J}$ is minimized subsequently on search spaces V_k which stem from the decomposition of the domain space $\mathbb{R}^n$ according to block-separability of $\mathcal{J}$. In our notation here, we would have $V_k \coloneqq \mathbb{R}^{n_k}$ and thereby simply minimize $\mathcal{J}$ with respect to the corresponding components of its arguments, adding up the respective solutions to the so-called *intermediate iterate* $u^{\nu+1/2}$. The approximate equality to the minimizer in (A.1.3) suggests that even inexact solution of the corresponding minimization problem is sufficient for satisfying convergence results. Details on this technically delicate matter and possible definitions of suitable solvers for (A.1.3) have been discussed in [33, Section 5].

Algorithm 13: Update Step Computation for the TNNMG Method

Input: Current iterate u^ν
begin Non-Linear Pre-Smoothing
  Set $w^{\nu,0} \coloneqq u^\nu$;
  for $k = 1,\dots,M$ **do**
    Compute $w^{\nu,k} \in w^{\nu,k-1} + V_k$ as

$$w^{\nu,k} :\approx \operatorname*{arg\,min}_{v \in w^{\nu,k-1}+V_k} \mathcal{J}(v)\,; \tag{A.1.3}$$

  end
  Set $u^{\nu+1/2} \coloneqq w^{\nu,m}$;
end
begin Truncated Linear Correction
  Determine large subspace $W_\nu \subset \mathbb{R}^n$ such that $\mathcal{J}\big|_{u^{\nu+1/2}+W_\nu}$ is twice continuously differentiable near $u^{\nu+1/2}$;
  Compute $v^\nu \in W_\nu$ as

$$v_\nu :\approx -\big(\mathcal{J}''\big(u^{\nu+1/2}\big)\big|_{W_\nu \times W_\nu}\big)^{-1}\big(\mathcal{J}'\big(u^{\nu+1/2}\big)\big|_{W_\nu}\big)\,; \tag{A.1.4}$$

end
begin Post-Processing
  Compute the projection $\tilde{v}^\nu \coloneqq \Pi_{\mathrm{dom}\mathcal{J}-u^{\nu+1/2}}(v^\nu)$;
  Compute step length $\rho_\nu \in [0,\infty[$ such that $\mathcal{J}\big(u^{\nu+1/2}+\rho_\nu\tilde{v}^\nu\big) \le \mathcal{J}\big(u^{\nu+1/2}\big)$;
end
Output: Set $u^{\nu+1} \coloneqq u^{\nu+1/2} + \rho_\nu\tilde{v}^\nu$;

Afterwards, we compute the *truncated linear correction* which under adequate assumptions enables fast convergence of the method. It consists of an (inexact) Newton step in an iteration-dependent subspace W_ν of $\mathbb{R}^n$ on which the objective functional $\mathcal{J}$ is sufficiently regular. Apparently, depending on the problem, the practical construction of this subspace can be technical. We note here that the subspaces do not have to be chosen *optimally* for the convergence results from [33], i.e., we do not have to find the largest subspace possible such that $\mathcal{J}$ is twice continuously differentiable on

$$\big(u^{\nu+1/2}+W_\nu\big) \cap B_\varepsilon\big(u^{\nu+1/2}\big)$$

for some $\varepsilon > 0$ but merely need one that allows for a well-defined Newton problem in (A.1.4). If $\mathcal{J}_0$ from the splitting (A.1.1) is a C^2-functional, W_ν can be straight-forwardly defined using a product space of $W_{\nu,k} \subset \mathbb{R}^{n_k}$ on which the φ_k from the block-separability condition (A.1.2) are locally smooth. A possibility resulting from this point of view is disabling entire so-called *inactive blocks* where the φ_k are non-smooth and are thus not considered for the computation of the Newton correction. For functionals as considered across all of the numerical investigations in this manuscript, i.e., where the non-smoothness is a block-wise norm function, this procedure provides the optimal smooth subspace in the sense explained beforehand. More general examples and more involved constructions of suitable subspaces for the truncated linear correction are considered in [33, Section 6].

At last, simply adding the linear coarse grid correction v^ν to the intermediate iterate $u^{\nu+1/2}$ in order to obtain the new iterate might lead to infeasibility since the definition of the former is not aware of the domain of $\mathcal{J}$. For this reason, a suitable *post-processing* procedure is often necessary. In order to both ensure feasibility of the updated iterate and to avoid very small damping parameters which might lead to poor convergence, the theoretical update $u^{\nu+1/2} + v^\nu$ is first projected onto the domain of $\mathcal{J}$ and the ensuing update step is then scaled such that objective decrease is achieved. This stands in contrast to mere scaling of the Newton update where aforementioned small damping parameters might appear. Across our numerical investigations, however, the effective domain of the objective functional is never restricted since values equal to ∞ are not possible for our concrete definitions of the respective $\mathcal{J}$. As a consequence, post-processing procedures do not feature non-trivial projection steps and only consist of adequately scaling the linear correction from before such that objective decrease is achieved.

A.2 Projection Algorithm onto the Special Linear Group

The algorithm is inspired by [26] and starts with a quasi-projection step in the direction $\operatorname{cof}(A)$. Note that the extrinsic derivative $\nabla \det(\cdot) = \operatorname{cof}(\cdot)$ is Frobenius-orthogonal to the polynomial restriction $\det(\cdot) = 1$ of the manifold $\mathfrak{P}$. Therefore, the initial iterate is given by

$$\tilde{P}^0 \coloneqq A + \gamma_0 \operatorname{cof}(A) = A + \gamma_0 \det(A) A^{-T} ,$$

with $\gamma_0 \in \mathbb{R}$ such that $\det(\tilde{P}^0) = 1$, and hence $\tilde{P}^0 \in \mathfrak{P}$. From there, we compute a sequence $(\tilde{P}^i)_{i\in\mathbb{N}}$ in $\mathfrak{P}$ with the following procedure: For each $\tilde{P}^i \in \mathfrak{P}$, let $\tilde{Q}^i$ denote the closest point to A in the tangential space $T_{\tilde{P}^i}\mathfrak{P}$. Then, the next quasi-projection is done in direction $\tilde{Q}^i - A$. For the next iterate $\tilde{P}^{i+1}$, find a $\gamma^{i+1} \in \mathbb{R}$ such that

$$\tilde{P}^{i+1} \coloneqq A + \gamma^{i+1}(\tilde{Q}^i - A)$$

is again in $\mathfrak{P}$. Although we have no convergence proof at hand, we observe similar convergence properties as those in [26]. In practice, the algorithm stopped after at most four iterations in our simulations with the criterion $\left\|\tilde{P}^{i+1} - \tilde{P}^i\right\|_F \leq 10^{-8}$. For an efficient implementation, note that it is not necessary to compute the points $\tilde{Q}^i$ explicitly. Since the direction $\tilde{Q}^i - A$ is orthogonal to $T_{\tilde{P}^i}\mathfrak{P}$, we can directly use the direction $\operatorname{cof}(\tilde{P}^i)$ and compute the next iterate by

$$\tilde{P}^{i+1} \coloneqq A + \gamma^{i+1} \operatorname{cof}(\tilde{P}^i) .$$

The γ^{i+1} can be computed directly by solving a cubic equation and afterwards choosing the solution with the smallest absolute value.

A.3 Test Machine Specifications

All tests are executed single-threaded on a Intel(R) Core(TM) i5-8265U CPU with clock frequency fixed to 1600 Mhz in order to avoid overheating and to ensure comparability of all test runs. The test machine runs the current snapshot of Debian 12, including updates as of January 30, 2023. The C++ Codes are compiled with the flags -03 -DNDEBUG using the gcc compiler in version 12.2.0.

A.4 Data Availability Statement

The easiest way to access the data evaluated over the course of the numerical investigations of our method is to directly recompute the results under the configuration which we have used. Let us give a short instruction on how to download the respective C++-codes together with the libraries required for conducting the tests in a *docker* image:[20]

First, create a directory where you save the docker file available from the link

Dockerfile Link

under the name *Dockerfile.txt* .[21] Afterwards, open a terminal in that directory and use the command

sudo docker image build -t dissertation-dune-docker .

(including the dot at the end) in order to create a virtual machine in which you can reproduce the directory structure which we have used in order to conduct the numerical tests. When encountering an error message similar to

fatal: unable to access '...': Could not resolve host: ...

just restart the procedure by reusing the command above. After successfully having created the docker image, you can run it in *interactive terminal mode* by using the command

sudo docker run -it dissertation-dune-docker:latest

which lets you end up in the terminal view of the desired directory structure. The directories *dune-proxnewton-modelproblems* and *dune-plasticity* contain the files necessary in order to reproduce the computations from the respective sections of numerical results. The structure is straight-forward and directory names within *src/Benchmarks* make it easy to navigate to the test which has to be conducted. Having reached the desired directory, you can then use the executable *run.sh*-file in order to start the computation.

Apparently, also all of the code used for the computations can be found and examined within the docker image. Thus, you can also easily implement changes to the existing methods and compile them using the *make* command in the respective *build-release* directory.

After having completed your numerical research on Proximal Newton methods, you can log out from the docker image using *Ctrl.+d*.

[20] Obviously, this requires having installed *docker* which can be accessed by using *sudo apt install docker.io* .

[21] For printed versions: The full link is https://gitlab.mn.tu-dresden.de/jaap/dune-plasticity-snapshot-poetzl/-/blob/82551a1fcaa32a7092c6e4e8e05d959c5a6b7d01/Dockerfile.

List of Symbols

Rate-Independent Systems

Finite Strain Plasticity

z	Placeholder for dissipative variables	30
A	Placeholder for gradient contributions	30
W_{el}, W_{pl}	Elastic and plastic stored energy density	30
$\boldsymbol{y}_D$	Time-dependent Dirichlet data	30
$\mathcal{Y}$	Space of admissible deformations	31
$\mathcal{Z}$	Domain of the internal variable	31
$\mathcal{E}$	Stored energy functional	31
$\mathcal{D}$	Dissipation distance	31
p_{el}, p_{pl}, p_{hd}, p_{gr}	Coercivity/Lebesgue exponents	32
$\boldsymbol{y}^k$, $\mathbf{P}^k$	Time-incremental solutions at time t_k	34
$\mathrm{SL}(3)^+_{\mathrm{sym}}$	Symmetric, positive definite plastic range	37
$\delta\mathbf{P}^+_{\mathrm{sym}}$	Symmetric plastic increment mapping	39
$\delta\mathbf{B}$	Tangential plastic increment mapping	39

Differentials and Semi-Smoothness Theory

∂, ∂^y	Convex subdifferential (with respect to a variable)	4
∂_F, ∂^y_F	Fréchet subdifferential (with respect to a variable)	4
∂_B	Bouligand subdifferential	52
∂_G	Clarke's generalized Jacobian	52
$F'(x,s)$	Directional derivative of F at x in direction s	53
∂^*	Generalized differential for semi-smoothness	54
$\partial^{(2)}$	Generalized second order differential	84

Algorithmic Quantities

prox_g	Euclidean proximal operator of the function g	59
prox_g^H	Scaled euclidean proximal operator of the function g	69
$G_L^{f,g}$	Euclidean composite gradient mapping of the sum $f+g$	60
L_f	Lipschitz-constant of the derivative f'	66
M	Uniform bound on the second order bilinear forms	66
κ_1	Convexity contribution of the second order bilinear form	67
κ_2	Convexity contribution of the non-smooth part	67
Δx	Full update step	68
$\mathcal{P}_g^H$	Generalized scaled proximal operator of the function g	70
x_*	Stationary point or optimal solution	73
$\lambda_{x,\omega}$	Regularized second order model decrease functional	76
$\Delta x(\omega)$	Damped update step	76
$x_+(\omega)$	Updated iterate by damped step	76
ω	Regularization parameter	76
γ	Sufficient decrease parameter	77
$\gamma(x,\omega)$	Value of the decrease ratio function	97
$\tilde{\gamma}(x,\omega)$	Value of the alternative decrease ratio function	103
G_τ^Φ	Generalized composite gradient mapping of the sum Φ	115
$\hat{F}_{x,\omega}$	Modified quadratic model of F around $x \in X$	116
$\Delta s(\omega)$	Inexact candidate for the damped update step	120

η	Forcing term for the relative error criterion	120
$\lambda^{\mu}_{x,\hat{\omega}}$	Regularized subgradient model decrease	124
$\Delta x^{\mu}(\hat{\omega})$	Subgradient step	124
$\tilde{\omega}_{\max}$	Subgradient regularization bound	125
ρ	Safety parameter for regularization strategies	152
a_{red}, p_{red}	Actual and predicted reduction for regularization	152
$\text{Rad}_{\rho,\gamma,p}$	Radicand function for the controller strategy	155
r	Reduction quotient for model-based forcing	162
$\omega_{\max}$	Maximal value of the regularization parameter	164

Miscellaneous

$\mathcal{X}_C$	Characteristic function of the set C	12
ϕ^*	Convex conjugate of the function ϕ	12
$\operatorname{dom}\phi$	Domain of the function ϕ	17
o, O	Landau symbols for convergence rates	45
$\mathfrak{R}$	Riesz-isomorphism on a Hilbert space	66
$\mathcal{L}(X,Y)$	Space of linear mappings from X to Y	4
X^*	Dual space of a Banach space X	4
$\mathcal{L}^{(2)}(X,Y)$	Space of bounded Y-vector valued bilinear forms on X	84

List of Algorithms

List of Figures

List of Tables

Bibliography

[1] H.-D. Alber. "Materials with Memory". In: *Lecture Notes in Mathematics* 1682 (1998).

[2] H.-B. An, Z.-Y. Mo, and X.-P. Liu. "A choice of forcing terms in inexact Newton method". In: *Journal of Computational and Applied Mathematics* 200.1 (2007), pp. 47–60. DOI: 10.1016/j.cam.2005.12.030.

[3] A. Y. Aravkin, R. Baraldi, and D. Orban. "A Proximal Quasi-Newton Trust-Region Method for Nonsmooth Regularized Optimization". en. In: (2021). DOI: 10.13140/RG.2.2.18509.15845/1.

[4] A. Argyriou et al. "Efficient First Order Methods for Linear Composite Regularizers". In: *Preprint* (Apr. 2011). arXiv: 1104.1436 [cs.LG].

[5] J. M. Ball. "Convexity conditions and existence theorems in nonlinear elasticity". In: *Archive for Rational Mechanics and Analysis* 63.4 (1976), pp. 337–403. DOI: 10.1007/bf00279992.

[6] H. T. Banks, S. Hu, and Z. R. Kenz. "A Brief Review of Elasticity and Viscoelasticity for Solids". In: *Advances in Applied Mathematics and Mechanics* 3.1 (2011), pp. 1–51. DOI: 10.4208/aamm.10-m1030.

[7] A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017. ISBN: 9781611974980. URL: https://doi.org/10.1137/1.9781611974997.

[8] H. Berninger, R. Kornhuber, and O. Sander. "Fast and Robust Numerical Solution of the Richards Equation in Homogeneous Soil". In: *SIAM Journal on Numerical Analysis* 49.6 (2011), pp. 2576–2597. DOI: 10.1137/100782887.

[9] J. Bolte, A. Daniilidis, and A. Lewis. "Tame functions are semismooth". In: *Mathematical Programming* 117.1-2 (2007), pp. 5–19. DOI: 10.1007/s10107-007-0166-9.

[10] O. T. Bruhns. "The Multiplicative Decomposition of the Deformation Gradient in Plasticity-Origin and Limitations". In: *Advanced Structured Materials*. Springer International Publishing, 2015, pp. 37–66. DOI: 10.1007/978-3-319-19440-0_3.

[11] R. H. Byrd, J. Nocedal, and F. Oztoprak. "An inexact successive quadratic approximation method for L-1 regularized optimization". In: *Mathematical Programming* 157.2 (2015), pp. 375–396. URL: https://doi.org/10.1007/s10107-015-0941-y.

[12] D.-Q. Chen, Y. Zhou, and L.-J. Song. "Fixed point algorithm based on adapted metric method for convex minimization problem with application to image deblurring". In: *Advances in Computational Mathematics* 42.6 (2016), pp. 1287–1310. URL: https://doi.org/10.1007/s10444-016-9462-3.

[13] P. Chen, J. Huang, and X. Zhang. "A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions". In: *Fixed Point Theory and Applications* 2016.1 (2016). URL: https://doi.org/10.1186/s13663-016-0543-2.

[14] P. G. Ciarlet. *Mathematical Elasticity. Three-Dimensional Elasticity (Studies in Mathematics and Its Applications, Vol 20)*. North Holland, 1994, p. 451. ISBN: 9780444817761.

[15] F. H. Clarke. *Optimization and Nonsmooth Analysis (Classics in Applied Mathematics)*. Society for Industrial Mathematics, 1987, p. 320. ISBN: 9780898712568.

[16] C. Clason and B. Jin. "A Semismooth Newton Method for Nonlinear Parameter Identification Problems with Impulsive Noise". In: *SIAM Journal on Imaging Sciences* 5.2 (2012), pp. 505–536. DOI: 10.1137/110826187.

[17] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000. DOI: 10.1137/1.9780898719857.

[18] Y. F. Dafalias. "Plastic spin: necessity or redundancy?" In: *International Journal of Plasticity* 14.9 (1998), pp. 909–931. DOI: 10.1016/s0749-6419(98)00036-9.

[19] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. "Inexact Newton Methods". In: *SIAM Journal on Numerical Analysis* 19.2 (1982), pp. 400–408. DOI: 10.1137/0719025.

[20] A. P. Dempster. "Covariance Selection". In: *Biometrics* 28.1 (1972), p. 157. DOI: 10.2307/2528966.

[21] P. Deuflhard. *Numerische Mathematik 2 gewöhnliche Differentialgleichungen. gewöhnliche Differentialgleichungen*. De Gruyter, 2013, p. 499. ISBN: 9783110316339.

[22] Q. T. Dinh, A. Kyrillidis, and V. Cevher. "A proximal Newton framework for composite minimization: Graph learning without Cholesky decompositions and matrix inversions". In: *Proceedings of the 30th International Conference on Machine Learning* (Jan. 8, 2013). arXiv: 1301.1459 [stat.ML].

[23] C. Eckart. "The Thermodynamics of Irreversible Processes. IV. The Theory of Elasticity and Anelasticity". In: *Physical Review* 73.4 (1948), pp. 373–382. DOI: 10.1103/physrev.73.373.

[24] S. C. Eisenstat and H. F. Walker. "Choosing the Forcing Terms in an Inexact Newton Method". In: *SIAM Journal on Scientific Computing* 17.1 (1996), pp. 16–32. DOI: 10.1137/0917003.

[25] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. Society for Industrial and Applied Mathematics, 1999. DOI: 10.1137/1.9781611971088.

[26] V. Elser. "Matrix product constraints by projection methods". In: *Journal of Global Optimization* 68.2 (2016), pp. 329–355. DOI: 10.1007/s10898-016-0466-9.

[27] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010, p. 749. ISBN: 9780821849743.

[28] H. Federer. *Geometric measure theory*. Springer, 1996, p. 676. ISBN: 3540606564.

[29] K. Fountoulakis and R. Tappenden. "A flexible coordinate descent method". In: *Computational Optimization and Applications* 70.2 (2018), pp. 351–394. URL: https://doi.org/10.1007/s10589-018-9984-3.

[30] G. Francfort and A. Mielke. "Existence results for a class of rate-independent material models with nonconvex elastic energies". In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 2006.595 (2006). DOI: 10.1515/crelle.2006.044.

[31] M. Fukushima and H. Mine. "A generalized proximal point algorithm for certain non-convex minimization problems". In: *International Journal of Systems Science* 12.8 (1981), pp. 989–1000. URL: https://doi.org/10.1080/00207728108963798.

[32] H. Ghanbari and K. Scheinberg. "Proximal quasi-Newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates". In: *Computational Optimization and Applications* 69.3 (2017), pp. 597–627. URL: https://doi.org/10.1007/s10589-017-9964-z.

[33] C. Gräser and O. Sander. "Truncated nonsmooth Newton multigrid methods for block-separable minimization problems". In: *IMA Journal of Numerical Analysis* 39.1 (2018), pp. 454–481. URL: https://doi.org/10.1093/imanum/dry073.

[34] M. E. Gurtin. *An introduction to continuum mechanics*. Academic Press, 1981, p. 265. ISBN: 0123097509.

[35] H. Hahn. "Leçons sur l'intégration et la recherche des fonctions primitives". In: *Monatshefte für Mathematik und Physik* 15.1 (1904), A46–A47. DOI: 10.1007/bf01692367.

[36] W. Han and B. Reddy. *Plasticity: Mathematical Theory and Numerical Analysis*. Springer, 1999. ISBN: 978-1-4614-5939-2.

[37] H. Hardering and O. Sander. "Geometric Finite Elements". In: *Handbook of Variational Methods for Nonlinear Geometric Data*. Springer International Publishing, 2020, pp. 3–49. DOI: 10.1007/978-3-030-31351-7_1.

[38] N. J. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, 2008. DOI: 10.1137/1.9780898717778.

[39] R. Hill. *The mathematical theory of plasticity*. Clarendon Press, 1998, p. 355. ISBN: 0198503679.

[40] M. Hintermüller and M. Ulbrich. "A mesh-independence result for semismooth Newton methods". In: *Math. Program.* 101.1, Ser. B (2004), pp. 151–184. ISSN: 0025-5610. URL: https://doi.org/10.1007/s10107-004-0540-9.

[41] M. Hintermüller, K. Ito, and K. Kunisch. "The Primal-Dual Active Set Strategy as a Semismooth Newton Method". In: *SIAM Journal on Optimization* 13.3 (2002), pp. 865–888. URL: https://doi.org/10.1137/s1052623401383558.

[42] M. Hinze. "A Variational Discretization Concept in Control Constrained Optimization: The Linear-Quadratic Case". In: *Computational Optimization and Applications* 30.1 (2005), pp. 45–61. DOI: 10.1007/s10589-005-4559-5.

[43] P. Jaap. "Efficient and Globally Convergent Minimization Algorithms for Small- and Finite-Strain Plasticity Problems". PhD thesis. TU Dresden, Feb. 2023.

[44] C. Kanzow and T. Lechner. "Globalized inexact proximal Newton-type methods for non-convex composite functions". In: *Computational Optimization and Applications* (2020). URL: https://doi.org/10.1007/s10589-020-00243-6.

[45] A. S. Khan. *Continuum theory of plasticity*. Wiley, 1995, p. 421. ISBN: 0471310433.

[46] W. T. Koiter. "Stress-strain relations, uniqueness and variational theorems for elastic-plastic materials with a singular yield surface". In: *Quarterly of Applied Mathematics* 11.3 (1953), pp. 350–354. DOI: 10.1090/qam/59769.

[47] K. KONDO. "A Proposal of a New Theory concerning the Yielding of Materials based on Riemannian Geometry, I". In: *Journal of the Society of Applied Mechanics of Japan* 2.11 (1949), pp. 123–128. DOI: 10.2322/jjsass1948.2.123.

[48] R. Kornhuber and R. Krause. "Robust Multigrid Methods for Vector-valued Allen–Cahn Equations with Logarithmic Free Energy". In: *Computing and Visualization in Science* 9.2 (2006), pp. 103–116. DOI: 10.1007/s00791-006-0020-2.

[49] A. M. Kosevich et al. *Theory of Elasticity, Third Edition. Volume 7 (Theoretical Physics, Vol 7)*. Butterworth-Heinemann, 1986, p. 195. ISBN: 9780750626330.

[50] R. Krause and O. Sander. "Fast Solving of Contact Problems on Complicated Geometries". In: *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, 2005, pp. 495–502. DOI: 10.1007/3-540-26825-1_51.

[51] A. Y. Kruger. "On Fréchet Subdifferentials". In: *Journal of Mathematical Sciences* 116.3 (2003), pp. 3325–3358. URL: https://doi.org/10.1023/a:1023673105317.

[52] B. Kummer. "Newton's Method Based on Generalized Derivatives for Nonsmooth Functions: Convergence Analysis". In: *Lecture Notes in Economics and Mathematical Systems*. Springer Berlin Heidelberg, 1992, pp. 171–194. DOI: 10.1007/978-3-642-51682-5_12.

[53] C. pei Lee and S. J. Wright. "Inexact Successive quadratic approximation for regularized optimization". In: *Computational Optimization and Applications* 72.3 (2019), pp. 641–674. URL: https://doi.org/10.1007/s10589-019-00059-z.

[54] E. H. Lee. "Elastic-Plastic Deformation at Finite Strains". In: *Journal of Applied Mechanics* 36.1 (1969), pp. 1–6. DOI: 10.1115/1.3564580.

[55] J. D. Lee, Y. Sun, and M. A. Saunders. "Proximal Newton-Type Methods for Minimizing Composite Functions". In: *SIAM Journal on Optimization* 24.3 (2014), pp. 1420–1443. URL: https://doi.org/10.1137/130921428.

[56] J. Li, M. S. Andersen, and L. Vandenberghe. "Inexact proximal Newton methods for self-concordant functions". In: *Mathematical Methods of Operations Research* 85.1 (2016), pp. 19–41. URL: https://doi.org/10.1007/s00186-016-0566-9.

[57] Q. Li et al. "Multi-step fixed-point proximity algorithms for solving a class of optimization problems arising from image processing". In: *Advances in Computational Mathematics* 41.2 (2014), pp. 387–422. URL: https://doi.org/10.1007/s10444-014-9363-2.

[58] L. Lubkoll, A. Schiela, and M. Weiser. "An affine covariant composite step method for optimization with PDEs as equality constraints". In: *Optimization Methods and Software* 32.5 (2016), pp. 1132–1161. DOI: 10.1080/10556788.2016.1241783.

[59] J. Lubliner. "On the thermodynamic foundations of non-linear solid mechanics". In: *International Journal of Non-Linear Mechanics* 7.3 (1972), pp. 237–254. DOI: 10.1016/0020-7462(72)90048-0.

[60] A. Mainik and A. Mielke. "Global Existence for Rate-Independent Gradient Plasticity at Finite Strain". In: *Journal of Nonlinear Science* 19.3 (2008), pp. 221–248. DOI: 10.1007/s00332-008-9033-y.

[61] G. D. Maso, G. A. Francfort, and R. Toader. "Quasistatic Crack Growth in Nonlinear Elasticity". In: *Archive for Rational Mechanics and Analysis* 176.2 (2005), pp. 165–225. DOI: 10.1007/s00205-004-0351-4.

[62] A. Mielke. "Evolution of Rate-Independent Systems". In: *Handbook of Differential Equations Evolutionary Equations*. Elsevier, 2005, pp. 461–559. DOI: 10.1016/s1874-5717(06)80009-5.

[63] A. Mielke. "Finite elastoplasticity Lie groups and geodesics on SL (d)". In: *Geometry, mechanics, and dynamics*. Springer, 2002, pp. 61–90. DOI: 10.1007/0-387-21791-6_2.

[64] A. Mielke and T. Roubiček. "Numerical approaches to rate-independent processes and applications in inelasticity". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 43.3 (2009), pp. 399–428. DOI: 10.1051/m2an/2009009.

[65] A. Mielke and T. Roubiček. "Rate-independent elastoplasticity at finite strains and its numerical approximation". In: *Mathematical Models and Methods in Applied Sciences* 26.12 (2016), pp. 2203–2236. DOI: 10.1142/s0218202516500512.

[66] A. Mielke and T. Roubíček. *Rate-Independent Systems*. Springer New York, 2015. DOI: 10.1007/978-1-4939-2706-7. URL: https://doi.org/10.1007/978-1-4939-2706-7.

[67] A. Mielke, F. Theil, and V. I. Levitas. "A Variational Formulation of Rate-Independent Phase Transformations Using an Extremum Principle". In: *Archive for Rational Mechanics and Analysis* 162.2 (2002), pp. 137–177. DOI: 10.1007/s002050200194.

[68] R. Mifflin. "Semismooth and Semiconvex Functions in Constrained Optimization". In: *SIAM Journal on Control and Optimization* 15.6 (1977), pp. 959–972. DOI: 10.1137/0315061.

[69] A. Milzarek. "Numerical methods and second order theory for nonsmooth problems". PhD thesis. TU München, 2016.

[70] A. Milzarek and M. Ulbrich. "A Semismooth Newton Method with Multidimensional Filter Globalization for l_1-Optimization". In: *SIAM Journal on Optimization* 24.1 (2014), pp. 298–333. URL: https://doi.org/10.1137/120892167.

[71] R. von Mises. "Mechanik der festen Körper in plastisch-deformablen Zustand". In: *Nachrichten der königlichen Gesellschaft der Wissenschaften Göttingen, Math.-phys. Klasse 4* (1913).

[72] M. Mooney. "A Theory of Large Elastic Deformation". In: *Journal of Applied Physics* 11.9 (1940), pp. 582–592. DOI: 10.1063/1.1712836.

[73] J. J. Moreau. "On Unilateral Constraints, Friction and Plasticity". In: *New Variational Techniques in Mathematical Physics*. Springer Berlin Heidelberg, 2011, pp. 171–322. DOI: 10.1007/978-3-642-10960-7_7.

[74] I. P. Natanson and L. F. Boron. *Theory of Functions of a Real Variable*. Dover Publications, Incorporated, 2016, p. 560. ISBN: 9780486806433.

[75] S. Nemat-Nasser. "On finite deformation elasto-plasticity". In: *International Journal of Solids and Structures* 18.10 (1982), pp. 857–872. DOI: 10.1016/0020-7683(82)90070-1.

[76] R. W. Ogden. "Large deformation isotropic elasticity – on the correlation of theory and experiment for incompressible rubberlike solids". In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 326.1567 (1972), pp. 565–584. DOI: 10.1098/rspa.1972.0026.

[77] L. Onsager. "Reciprocal Relations in Irreversible Processes. I." In: *Physical Review* 37.4 (1931), pp. 405–426. DOI: 10.1103/physrev.37.405.

[78] M. Ortiz and E. Repetto. "Nonconvex energy minimization and dislocation structures in ductile single crystals". In: *Journal of the Mechanics and Physics of Solids* 47.2 (1999), pp. 397–462. DOI: 10.1016/s0022-5096(97)00096-3.

[79] M. Ortiz and L. Stainier. "The variational formulation of viscoplastic constitutive updates". In: *Computer Methods in Applied Mechanics and Engineering* 171.3-4 (1999), pp. 419–444. DOI: 10.1016/s0045-7825(98)00219-9.

[80] J.-S. Pang and L. Qi. "Nonsmooth Equations: Motivation and Algorithms". In: *SIAM Journal on Optimization* 3.3 (1993), pp. 443–465. DOI: 10.1137/0803021.

[81] E. Pipping, O. Sander, and R. Kornhuber. "Variational formulation of rate- and state-dependent friction problems". In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 95.4 (2013), pp. 377–395. DOI: 10.1002/zamm.201300062.

[82] W. Prager and P. G. Hodge. *Theory of perfectly plastic solids*. New York: John Wiley, 1951. ISBN: 2566893207448.

[83] B. Pötzl, A. Schiela, and P. Jaap. "Inexact Proximal Newton methods in Hilbert spaces". In: (Apr. 2022). DOI: 10.48550/ARXIV.2204.12168. arXiv: 2204.12168 [math.OC].

[84] B. Pötzl, A. Schiela, and P. Jaap. "Second order semi-smooth Proximal Newton methods in Hilbert spaces". In: *Computational Optimization and Applications* 82.2 (2022), pp. 465–498. DOI: 10.1007/s10589-022-00369-9.

[85] L. Qi. "Convergence Analysis of Some Algorithms for Solving Nonsmooth Equations". In: *Mathematics of Operations Research* 18.1 (1993), pp. 227–244. DOI: 10.1287/moor.18.1.227.

[86] L. Qi and J. Sun. "A nonsmooth version of Newton's method". In: *Mathematical Programming* 58.1-3 (1993), pp. 353–367. DOI: 10.1007/bf01581275.

[87] E. Ramm et al. *Error-controlled Adaptive Finite Elements in Solid Mechanics*. Wiley, 2001, p. 500. ISBN: 9780471496502.

[88] R. S. Rivlin. "Large elastic deformations of isotropic materials IV. further developments of the general theory". In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 241.835 (1948), pp. 379–397. DOI: 10.1098/rsta.1948.0024.

[89] S. M. Robinson. "Newton's method for a class of nonsmooth functions". In: *Set-Valued Analysis* 2.1-2 (1994), pp. 291–305. DOI: 10.1007/bf01027107.

[90] R. T. Rockafellar, M. Wets, and R. J. B. Wets. *Variational Analysis*. Springer London, Limited, 2009. ISBN: 9783642024313.

[91] S. Sadik and A. Yavari. "On the origins of the idea of the multiplicative decomposition of the deformation gradient". In: *Mathematics and Mechanics of Solids* 22.4 (2015), pp. 771–772. DOI: 10.1177/1081286515612280.

[92] O. Sander and P. Jaap. "Solving primal plasticity increment problems in the time of a single predictor–corrector iteration". In: *Computational Mechanics* 65.3 (Oct. 2019), pp. 663–685. DOI: 10.1007/s00466-019-01788-y. URL: https://doi.org/10.1007/s00466-019-01788-y.

[93] G. Scalet and F. Auricchio. "Computational Methods for Elastoplasticity: An Overview of Conventional and Less-Conventional Approaches". In: *Archives of Computational Methods in Engineering* 25.3 (2017), pp. 545–589. DOI: 10.1007/s11831-016-9208-x.

[94] M. Schaller, A. Schiela, and M. Stöcklein. "A Composite Step Method with Inexact Step Computations for PDE Constrained Optimization". In: *DFG Priority Programme 1962 Preprint* (2018). URL: https://spp1962.wias-berlin.de/preprints/098.pdf.

[95] E. Schechter. *Handbook of analysis and its foundations*. Academic Press, 1997, p. 883. ISBN: 0126227608.

[96] K. Scheinberg and X. Tang. "Practical inexact proximal quasi-Newton method with global complexity analysis". In: *Mathematical Programming* 160.1-2 (2016), pp. 495–529. URL: https://doi.org/10.1007/s10107-016-0997-3.

[97] A. Schiela. "A simplified approach to semismooth Newton methods in function space". In: *SIAM J. Optim.* 19.3 (2008), pp. 1417–1432. ISSN: 1052-6234. URL: https://doi.org/10.1137/060674375.

[98] J. C. Simo. *Computational inelasticity*. Springer, 1998, p. 392. ISBN: 0387975209.

[99] J. Simo. "A framework for finite strain elastoplasticity based on maximum plastic dissipation and the multiplicative decomposition: Part I. Continuum formulation". In: *Computer Methods in Applied Mechanics and Engineering* 66.2 (1988), pp. 199–219. DOI: 10.1016/0045-7825(88)90076-x.

[100] J. Simo and M. Ortiz. "A unified approach to finite deformation elastoplastic analysis based on the use of hyperelastic constitutive equations". In: *Computer Methods in Applied Mechanics and Engineering* 49.2 (1985), pp. 221–245. DOI: 10.1016/0045-7825(85)90061-1.

[101] W. S. Slaughter. *Linearized theory of elasticity*. Birkhäuser, 2002, p. 543. ISBN: 0817641173.

[102] L. Stella, A. Themelis, and P. Patrinos. "Forward–backward quasi-Newton methods for nonsmooth optimization problems". In: *Computational Optimization and Applications* 67.3 (2017), pp. 443–487. URL: https://doi.org/10.1007/s10589-017-9912-y.

[103] R. Temam and L. S. Orde. *Mathematical Problems in Plasticity*. Dover Publications, Incorporated, 2019, p. 368. ISBN: 9780486828275.

[104] Q. Tran-Dinh, Y.-H. Li, and V. Cevher. "Composite Convex Minimization Involving Self-concordant-Like Cost Functions". In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, 2015, pp. 155–168. URL: https://doi.org/10.1007/978-3-319-18161-5_14.

[105] H. Tresca. "Mémoire sur l'écoulement des corps solides". In: *Mém. prés. de Paris 18, 733-799* (1868).

[106] F. Tröltzsch. *Optimal control of partial differential equations*. Vol. 112. Graduate Studies in Mathematics. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels. American Mathematical Society, Providence, RI, 2010, pp. xvi+399. ISBN: 978-0-8218-4904-0. URL: https://doi.org/10.1090/gsm/112.

[107] P. Tseng and S. Yun. "A coordinate gradient descent method for nonsmooth separable minimization". In: *Mathematical Programming* 117.1-2 (2007), pp. 387–423. URL: https://doi.org/10.1007/s10107-007-0170-0.

[108] M. Ulbrich. *Nichtlineare Optimierung*. Birkhäuser Verlag, 2011, p. 150. ISBN: 9783034601429.

[109] M. Ulbrich. *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Habilitation Thesis. 2002.

[110] M. Ulbrich. "Semismooth Newton Methods for Operator Equations in Function Spaces". In: *SIAM Journal on Optimization* 13.3 (2002), pp. 805–841. DOI: 10.1137/s1052623400371569.

[111] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Society for Industrial and Applied Mathematics, 2011. URL: https://doi.org/10.1137/1.9781611970692.

[112] M. Ulbrich. *Semismooth Newton Methods in Function Space - Theoretical Foundations and Applications*. 42nd Woudschoten Conference. Oct. 2017.

[113] P. Urysohn. "Zum Metrisationsproblem". In: *Mathematische Annalen* 94.1 (1925), pp. 309–315. DOI: 10.1007/bf01208661.

[114] A. Walther and A. Griewank. "Getting Started with ADOL-C". In: *Combinatorial Scientific Computing*. Chapman and Hall/CRC, 2012, pp. 181–202. URL: https://doi.org/10.1201/b11644-8.

[115] M. Weiser, P. Deuflhard, and B. Erdmann. "Affine conjugate adaptive Newton methods for nonlinear elastomechanics". In: *Optimization Methods and Software* 22.3 (2007), pp. 413–431. DOI: 10.1080/10556780600605129.

[116] H. Xu. "Set-valued approximations and Newton's methods". In: *Mathematical Programming* 84.2 (1999), pp. 401–420. DOI: 10.1007/s101070050028.

[117] E. Zeidler. *Nonlinear Functional Analysis and Its Applications: Part 2 B. Nonlinear Monotone Operators (Zeidler, Eberhard//Nonlinear Functional Analysis and Its Applications)*. Springer, 1989, p. 756. ISBN: 9780387971674.

[118] H. Ziegler. "An attempt to generalize Onsager's principle, and its significance for rheological problems". In: *ZAMP Zeitschrift für Angewandte Mathematik und Physik* 9.5-6 (1958), pp. 748–763. DOI: 10.1007/bf02424793.

[119] H. Ziegler and C. Wehrli. "The Derivation of Constitutive Relations from the Free Energy and the Dissipation Function". In: *Advances in Applied Mechanics*. Elsevier, 1987, pp. 183–238. DOI: 10.1016/s0065-2156(08)70278-3.

[120] W. P. Ziemer. *Weakly Differentiable Functions*. Springer New York, 1989. DOI: 10.1007/978-1-4612-1015-3.

Publications

[83] B. Pötzl, A. Schiela, and P. Jaap. "Inexact Proximal Newton methods in Hilbert spaces". In: (Apr. 2022). DOI: 10.48550/ARXIV.2204.12168. arXiv: 2204.12168 [math.OC].

[84] B. Pötzl, A. Schiela, and P. Jaap. "Second order semi-smooth Proximal Newton methods in Hilbert spaces". In: *Computational Optimization and Applications* 82.2 (2022), pp. 465–498. DOI: 10.1007/s10589-022-00369-9.

Several parts of this thesis are contained in the publications above. The deliberations on dual scaled proximal mappings and ensuing algorithmic developments in Section 3.2 have first been considered in [84]. The latter work also in parts contains the introduction to the notion of second order semi-smoothness which we have elaborated on in a more detailed fashion here in Section 3.2.4. The alternative sufficient decrease criterion for numerical robustness close to optimal solutions from Section 3.2.6 has not been a part of this publication.

The submitted preprint [83] considers the inexact computation of update steps and thereby serves as a basis for the elaborations of Section 4.1. We have significantly extended the numerical investigations conducted for the influence of inexactness for the present manuscript. Also in that regard, the consideration of the alternative sufficient decrease criterion, cf. Section 4.1.5, crucially improved the results from the standpoint of both convergence analysis and numerical investigations for the version in this manuscript.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe.

Weiterhin erkläre ich, dass ich die Hilfe von gewerblichen Promotionsberatern bzw. -vermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe, noch künftig in Anspruch nehmen werde.

Zusätzlich erkläre ich hiermit, dass ich keinerlei früheren Promotionsversuche unternommen habe.

Bayreuth, den

Bastian Pötzl