JULIUS-MAXIMILIANS-UNIVERSITÄT WÜRZBURG
Fakultät für Mathematik und Informatik

# Proximal Methods for Nonconvex Composite Optimization Problems

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg

vorgelegt von

**Theresa Lechner**

Würzburg, 2022

| | |
|---|---|
| 1. Gutachter: | Prof. Dr. Christian Kanzow<br>Julius-Maximilians-Universität Würzburg |
| 2. Gutachter: | Prof. Dr. Peter Ochs<br>Eberhard Karls Universität Tübingen |

Theresa Lechner

# Proximal Methods for Nonconvex Composite Optimization Problems

Dissertation

# Abstract

Optimization problems with composite functions deal with the minimization of the sum of a smooth function and a convex nonsmooth function. In this thesis several numerical methods for solving such problems in finite-dimensional spaces are discussed, which are based on proximity operators.

After some basic results from convex and nonsmooth analysis are summarized, a first-order method, the proximal gradient method, is presented and its convergence properties are discussed in detail. Known results from the literature are summarized and supplemented by additional ones. Subsequently, the main part of the thesis is the derivation of two methods which, in addition, make use of second-order information and are based on proximal Newton and proximal quasi-Newton methods, respectively. The difference between the two methods is that the first one uses a classical line search, while the second one uses a regularization parameter instead. Both techniques lead to the advantage that, in contrast to many similar methods, in the respective detailed convergence analysis global convergence to stationary points can be proved without any restricting precondition. Furthermore, comprehensive results show the local convergence properties as well as convergence rates of these algorithms, which are based on rather weak assumptions. Also a method for the solution of the arising proximal subproblems is investigated.

In addition, the thesis contains an extensive collection of application examples and a detailed discussion of the related numerical results.

# Zusammenfassung

In Optimierungsproblemen mit zusammengesetzten Funktionen wird die Summe aus einer glatten und einer konvexen, nicht glatten Funktion minimiert. Die vorliegende Arbeit behandelt mehrere numerische Verfahren zur Lösung solcher Probleme in endlich-dimensionalen Räumen, welche auf *Proximity Operatoren* basieren.

Nach der Zusammenfassung einiger grundlegender Resultate aus der konvexen und nichtglatten Analysis wird ein Verfahren erster Ordnung, das *Proximal-Gradienten-Verfahren*, vorgestellt und dessen Konvergenzeigenschaften ausführlich behandelt. Bekannte Resultate aus der Literatur werden dabei zusammengefasst und durch weitere Ergebnisse ergänzt. Im Anschluss werden im Hauptteil der Arbeit zwei Verfahren hergeleitet, die zusätzlich Informationen zweiter Ordnung nutzen und auf *Proximal-Newton-* beziehungsweise *Proximal-Quasi-Newton-Verfahren* beruhen. Der Unterschied zwischen beiden Verfahren liegt darin, dass bei ersterem eine klassische Schrittweitensuche verwendet wird, während das zweite stattdessen einen Regularisierungsparameter nutzt. Beide Techniken führen dazu, dass im Gegensatz zu vielen verwandten Verfahren in der jeweils ausführlichen Konvergenzanalyse die globale Konvergenz zu stationären Punkten ohne weitere einschränkende Voraussetzungen bewiesen werden kann. Ferner zeigen umfassende Resultate die lokalen Konvergenzeigenschaften sowie Konvergenzraten der Algorithmen auf, welche auf lediglich schwachen Annahmen beruhen. Ein Verfahren zur Lösung auftretender *Proximal-Teilprobleme* ist ebenfalls Bestandteil dieser Arbeit.

Die Dissertation beinhaltet zudem eine umfangreiche Sammlung von Anwendungsbeispielen und zugehörigen numerischen Ergebnissen.

# Acknowledgement

> "Which is more important," asked Big Panda,
> "the journey or the destination?"
> "The company." said Tiny Dragon.
>
> *James Norbury* [125]

Remarkably, it is a pleasure for me to write this section as it reflects the extensive support and invaluable assistance I have received from many sides during my research for the development of this thesis in the last years.

First and foremost, I have to thank my supervisor, Prof. Christian Kanzow. Through his invaluable expertise and indispensable guidance, he has made this work possible in the first place. I also thank him for the countless mathematical and formal tools he shared with me, inspiring hints and occasionally good advice.

Furthermore, I would like to thank my former colleagues Dr. Eike Börgens, Tobias Geiger, Dr. Veronika Karl, and Carolin Natemeyer for the wonderful collaboration and inspiring discussions. Besides the scientific exchange, I especially appreciated joint activities and the motivating conversations. I am also grateful to my current colleagues Xiaoxi Jia, Tanja Neder, Simeon vom Dahl, and Felix Weiß for the joint work. In particular, I thank Dr. Eike Börgens, Carolin Natemeyer and Tanja Neder for proofreading this thesis.

I would like to express my gratitude to numerous further members of the Institute of Mathematics at the University of Würzburg, since they have shown me that science is not only about proving theorems and writing papers. I thank them for numerous informative conversations, exchanges on a wide variety of topics, and support with formal questions and bureaucratic issues.

Even though they were not directly involved in the creation of this thesis, I must express my great gratitude to Albrecht Kliem and Annett Rauch-Weise. Both teachers showed me the magic of mathematics during my school years, encouraged my inquisitiveness and made it possible for me to participate in numerous mathematical workshops as well as to start my first scientific work. For that, I am infinitely grateful to them even today.

Finally, I would like to thank all my friends, old and new, as well as my family, who have influenced the last few years. They have tolerated and put up with me when I spent the whole day brooding over theories and was too tired to do anything after work. At the same time, they motivated me when I lacked motivation, supported and distracted me when I was stuck within my thoughts, and showed me again and again that the beauty of life is much more than mathematics.

# Contents

# List of Symbols and Abbreviations

### Sets

| | |
|---|---|
| $\emptyset$ | empty set |
| $\mathbb{R}$ | set of real numbers |
| $\mathbb{R}_+, \mathbb{R}_{++}$ | set of nonnegative and positive real numbers |
| $\overline{\mathbb{R}}$ | set of extended real numbers, i.e. $\mathbb{R} \cup \{+\infty\}$ |
| $\mathbb{R}^n$ | space of $n$-dimensional real vectors |
| $\mathbb{N}$ | set of positive integers |
| $\mathbb{N}_0$ | set of nonnegative integers |
| $B_\varepsilon(x)$ | open ball with radius $\varepsilon > 0$ around $x$ w.r.t. the Euclidean norm |
| $\overline{S}$ | closure of the set $S$ |
| $\mathrm{conv}(S)$ | convex hull of the set $S$ |
| $\mathrm{int}(S)$ | interior of the set $S$ |

### Matrices and Vectors

| | |
|---|---|
| $I$ | identity matrix in $\mathbb{R}^{n\times n}$ |
| $\mathbb{S}^n$ | set of symmetric, real $n \times n$ matrices |
| $\mathbb{S}^n_+$ | set of symmetric, real, positive semidefinite $n \times n$ matrices |
| $\mathbb{S}^n_{++}$ | set of symmetric, real, positive definite $n \times n$ matrices |
| $\lambda_{\min}(M), \lambda_{\max}(M)$ | smallest and largest eigenvalue of a symmetric matrix $M$ |
| $M_{[\mathcal{I}\mathcal{J}]}$ | submatrix of a matrix $M \in \mathbb{R}^{m\times n}$ w.r.t. the index sets<br>$\mathcal{I} \subset \{1,\dots,m\}$, $\mathcal{J} \subset \{1,\dots,n\}$ |
| $x_{[\mathcal{I}]}$ | subvector of a vector $x \in \mathbb{R}^n$ w.r.t. the index set $\mathcal{I} \subset \{1,\dots,n\}$ |
| $x_i$ | $i$-th component of the vector $x$ |
| $M_{[i\cdot]}, M_{[\cdot j]}$ | $i$-th row and $j$-th column of the matrix $M$ |
| $\mathrm{diag}(x)$ | diagonal matrix with entries $\mathrm{diag}(x)_{[ii]} = x_i$ for $i = 1,\dots,n$ |
| $M^\dagger$ | (Moore-Penrose) pseudoinverse of the matrix $M$ |

### Operations on vectors and matrices

| | |
|---|---|
| $\succeq$, $\preceq$ | partial ordering in $\mathbb{S}^n$, i.e. $A \succeq B \;\Leftrightarrow B \preceq A \Leftrightarrow\; A - B \in \mathbb{S}^n_+$ |
| $\langle\cdot,\cdot\rangle$ | Euclidean inner product, i.e. $\langle x, y\rangle = x^T y$ |

| | |
|---|---|
| $\\|\cdot\\|, \\|\cdot\\|_2$ | Euclidean norm, i.e. $\\|x\\|^2 = \\|x\\|_2^2 = \langle x, x\rangle$ |
| $\\|\cdot\\|_1$ | $\ell_1$-norm, i.e. $\\|x\\|_1 = \sum_{i=1}^n \lvert x_i\rvert$ |
| $\\|\cdot\\|_\infty$ | $\ell_\infty$-norm, i.e. $\\|x\\|_\infty = \max_{i=1,\dots,n} \lvert x_i\rvert$ |
| $\langle\cdot,\cdot\rangle_H$ | inner product induced by $H \in \mathbb{S}^n_{++}$, i.e. $\langle x, y\rangle_H = x^T Hy$ |
| $\\|\cdot\\|_H$ | norm induced by $H \in \mathbb{S}^n_{++}$, i.e. $\\|x\\|_H^2 = \langle x, x\rangle_H$ |
| $\odot$ | component-wise product (Hadamard product),<br>i.e. $(x \odot y)_i = x_i \cdot y_i$ |
| $\operatorname{dist}(x, S)$ | distance between the point $x$ and the set $S$,<br>i.e. $\operatorname{dist}(x, S) = \inf_{z\in S} \\|z - x\\|$ |
| $\operatorname{sign}(x)$ | element-wise sign of the vector $x$ |

**Functions and operations on functions**

| | |
|---|---|
| $\operatorname{dom}\varphi$ | domain of the function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ |
| Id | identity mapping |
| $\operatorname{Im}(F)$ | image of the operator $F : \mathbb{R}^n \to \mathbb{R}^m$ |
| $\operatorname{lev}_{\leq\alpha}\varphi$ | lower level set of $\varphi$ at the level $\alpha \in \mathbb{R}$, i.e.<br>$\operatorname{lev}_{\leq\alpha}\varphi = \{x \in \mathbb{R}^n : \varphi(x) \leq \alpha\}$ |
| $\iota_S(\cdot)$ | indicator function of the set $S$ |
| $\varphi'(x; d)$ | directional derivative of $\varphi$ at $x$ in direction $d$ |
| $\partial\varphi(x)$ | convex subdifferential or Clarke's generalized gradient of $\varphi$ in $x$ |
| $\partial_B\varphi(x)$ | B-subdifferential/Bouligand-subdifferential of $\varphi$ in $x$ |
| $\partial_\varepsilon\varphi(x)$ | $\varepsilon$-subdifferential of $\varphi$ in $x$ |
| $\nabla f(x), \nabla^2 f(x)$ | gradient and Hessian of the function $f : \mathbb{R}^n \to \mathbb{R}$ in $x$ |

**Other**

| | |
|---|---|
| $\{x^k\}$ | sequence of points ($k = 1, 2, \dots$) |
| $\{x^k\}_\mathcal{K}$ | subsequence of $\{x^k\}$ with $k \in \mathcal{K}$ |
| $x^k = \mathcal{O}(y^k)$,<br>$x^k = o(y^k)$ | Landau-Symbols for sequences $\{x^k\} \subset \mathbb{R}^n$ and $\{y^k\} \subset \mathbb{R}$,<br>i.e. $\limsup_{k\to\infty} \frac{\\|x^k\\|}{\lvert y^k\rvert} < +\infty$, $\lim_{k\to\infty} \frac{x^k}{y^k} = 0$, resp. |
| $t_k \downarrow 0$ | convergence of the sequence $\{t_k\} \subset \mathbb{R}_+$ to 0 from above |

# Chapter 1

# Introduction

The subject of this thesis is the investigation of general optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \psi(x) := f(x) + \varphi(x), \tag{1.1}$$

where $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex, proper and lower semicontinuous mapping and $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth, possibly nonconvex function, as well as the detailed development and analysis of efficient numerical algorithms for their solution. The objective function in this problem is called *composite function* by several authors [40, 95, 121] and its minimization was formally introduced and investigated by Fukushima and Mine [66, 113] back in the 1980s.

## Motivation

In the above formulation, the objective function is not necessarily smooth nor convex, and, additionally, since $\varphi$ is an extended real valued function, even constrained problems can be formulated in this way. Therefore, the problem structure is sufficiently broad to cover a large variety of applications for real-life problems, for example in the areas of statistics, machine learning, compressed sensing, and signal processing. In result, the interest in problems in formulation (1.1) got a lot of attention from researchers, especially during the last decade, and innumerable numerical algorithms were developed for their solution. To show the generality and relevance of problem (1.1), some motivating examples of applications are provided.

**Example 1.1** (Compressed sensing)**.** Given a linear signal model by a matrix $A \in \mathbb{R}^{m\times n}$ and a (possibly noisy) observation vector $b \in \mathbb{R}^m$, the aim of many inverse problems is to reconstruct a sparse vector $x \in \mathbb{R}^n$ such that $Ax \approx b$. A natural formulation for the recovery problem with focus on sparsity [10] is

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \quad \text{such that } \|Ax - b\| \leq \varepsilon,$$

where $\varepsilon \geq 0$ is related to the occuring noise and $\|x\|_0$ denotes the $\ell_0$-norm counting the nonzero entries of $x$. Although it is called a norm, the $\ell_0$-norm is not a norm in the classical sense. Thus, whilst using the $\ell_0$-norm leads to a sparse solution, the $\ell_0$-norm optimization problem is difficult to solve due to its discontinuity and nonconvexity. Hence, the $\ell_0$-term is often replaced by the $\ell_1$-norm. Furthermore, instead of using the estimation of $Ax$ to $b$ as constraint, one can handle it as part of the objective function, which leads to the popular *Lasso* problem [151]

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1$$

for some $\lambda > 0$, a problem of type (1.1). This procedure works well, as long as the noise in the signal vector $b$ is Gaussian. However, for non-Gaussian noise the Euclidean norm in the quadratic term needs to be replaced. For example, if the error is obtained from Student's $t$-distribution [4, 112], we end up with the nonconvex problem

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^m \log\left(1 + \frac{(Ax-b)_i^2}{\nu}\right) + \lambda \|x\|_1$$

with degree of freedom $\nu > 0$. ♢

**Example 1.2** (Sparse logistic regression)**.** Logistic regression is used to separate feature vectors by a hyperplane, hence the problem setting is similar to the one of support vector machines. In logistic regression, feature vectors $a^1, \dots, a^m \in \mathbb{R}^n$ and corresponding binary labels $y_1, \dots, y_m \in \{-1, 1\}$ are given, which often represent some measured data. Since the probability distribution of the class label $y$ for a given feature vector $a \in \mathbb{R}^n$ and a logistic regression coefficient vector $x \in \mathbb{R}^n$ can be described by

$$p(y = 1 \mid a; y) = \frac{\exp(a^T x)}{1 + \exp(a^T x)},$$

the determination of this coefficient vector $x$ leads to the problem

$$\min_{x \in \mathbb{R}^n, \nu \in \mathbb{R}} \sum_{i=1}^m \log\big(1 + \exp(x^T a^i + \nu y_i)\big),$$

see [168]. In some cases, a drawback in considering this problem is overfitting [88], which is prevented by adding a regularization term. If the features are connected in several groups, this can be done with a group lasso term, i.e. for sets $\mathcal{I}_1, \dots, \mathcal{I}_s \subset \{1, \dots, n\}$ and some regularization parameter $\lambda > 0$, we get

$$\min_{x \in \mathbb{R}^n, \nu \in \mathbb{R}} \sum_{i=1}^m \log\big(1 + \exp(x^T a^i + \nu y_i)\big) + \lambda \sum_{j=1}^s \|x_{[\mathcal{I}_j]}\|,$$

which again has the form (1.1). If the groups $\mathcal{I}_1, \dots, \mathcal{I}_s$ form a partition of $\{1, \dots, n\}$, this is the classical group lasso, whereas we get the overlapping group lasso, if these sets are not pairwise disjoint [109]. ♢

**Example 1.3** (Constrained optimization)**.** A classical formulation of constrained optimization problems is

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to } h(x) = 0, \; g(x) \le 0$$

for $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^p$, which can equivalently be written as

$$\min_{x \in \mathbb{R}^n} f(x) + \iota_{\mathcal{X}}(x),$$

where $\mathcal{X} = \{x \in \mathbb{R}^n \mid g(x) \le 0, \; h(x) = 0\}$ and $\iota_{\mathcal{X}}$ denotes its indicator function. Assuming that $h$ is affine and the component functions $g_i$ for $i = 1, \dots, m$ are convex, the set $\mathcal{X}$ is closed and convex. Hence, if $f$ is smooth, we obtain a problem as formulated in (1.1), which therefore also covers a large class of constrained problems. ♢

There are countless further examples of applications whose problem setting can be affiliated to (1.1), for example (sparse) inverse covariance selection [78, 130, 144], blind deconvolution

[9,19,21] or nonnegative matrix factorization [94,102]. Further applications can be found in deep learning [54], data clustering [138] and dictionary learning [67,108]. Moreover, optimization problems involving composite functions are utilized in magnetic resonance imaging and tomography [18,160] or seismic tomography [124,149].
We note that this list contains only a small part of the applications and is by no means complete. Further examples can be found in the literature, e.g. [2,26,50].

## Proximal Methods

The rapid growth of convex optimization applications, in particular in signal processing and machine learning, has further increased the popularity of methods for solving (1.1). A significant part of these methods can be classified using the term *proximal methods*, which are considered in more detail below.
The foundation of proximal methods is the solution of a subproblem of the form

$$\arg\min_{d} \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + \varphi(x^k + d) \right\} \tag{1.2}$$

in each iteration step, where $H_k \in \mathbb{S}^n$ is usually a positive definite matrix. In that case, its solution $d^k$ is unique and characterized using the *proximity operator*

$$d^k = \operatorname{prox}_{\varphi}^{H_k} \left( x^k - H_k^{-1} \nabla f(x^k) \right). \tag{1.3}$$

In the latter representation it is apparent, why these methods are also called *forward backward methods* [16,32,161] (first, a forward step performs a gradient-step using information on $f$, and second, the backward step applies the proximity operator using information on $\varphi$), whereas formulation (1.2) motivates the term *successive quadratic approximation methods* [38].
In the simplest form, the next iterate is set to $x^{k+1} = x^k + d^k$, but due to a lack of convergence, this strategy is usually combined with some form of backtracking. There exist two main backtracking strategies. In the first one, convergence is guaranteed using backtracking by, roughly speaking, increasing $H_k$, e.g. by adding a positive multiple of the identity matrix, whereas for the second one we set $x^{k+1} = x^k + t_k d^k$ for an appropriate step length $t_k \in (0,1]$ (classical line search).
In the following, we provide an extensive overview over existing literature regarding proximal methods before devoting an introduction to the methods investigated in the subsequent chapters of this thesis.

**First-Order Methods** Among the most popular algorithms for solving problem (1.1) are the Iterative Shrinkage Thresholding Algorithm (ISTA) and its accelerated version FISTA [14,120]. Like all first-order proximal methods, in ISTA-type methods the matrix $H_k$ is chosen as a positive multiple of the identity, i.e. $H_k = L_k I$, where $L_k > 0$ is either constant during the algorithm or adapted to approximate a Lipschitz constant of $\nabla f$. The backtracking consists of choosing $L_k$ sufficiently large such that a suitable descent condition holds. Since the steps are computationally inexpensive in many cases or can even be solved analytically, this method has been the subject of intensive research over the last decade in numerous variants and with different acceleration techniques [73,79,80,129,143,146,161,164,171]. Various preliminaries on the functions $f$ and $\varphi$ are studied, but a common assumption and one of the drawbacks of these methods is that $\nabla f$ must in most cases be (globally) Lipschitz continuous to guarantee convergence.

In the class of heavy-ball or inertial proximal methods [27, 126–128], the approach of ISTA is combined with an extrapolation step. A further approach is to perform a classical line search and determine $t_k \in (0,1]$ such that a descent condition holds for $x^k + t_k d^k$ [17, 26, 156, 165, 168]. With this approach the convergence can be proved even without Lipschitz continuity. In recent years, also the Kurdyka-Łojasiewicz property is used to prove convergence rates. Since the interest in large scale problems increased in recent years, block coordinate and stochastic proximal methods have been developed [29, 62, 63, 93, 156]. For (block) separable functions $f$ and $\phi$, these are based on the idea of applying proximal steps only to a part of the indices of the iterates in each step. Finally, we mention some more first-order proximal methods [49, 50, 59] that do not fit into the mentioned categories.

**Variable Metric Proximal Methods** First-order methods are easy to implement and have well-studied convergence properties. However, they are well-known to have poor convergence rates, especially when high accuracy is required. This drawback may result in a large number of iterations to obtain an acceptable approximation of the solution. Variable metric techniques have therefore been proposed in recent literature [25, 28, 29, 31, 32, 46, 62–64, 93]. In variable metric proximal methods, the matrix $H_k$ and therefore the underlying metric may change in every iteration based on suitable criteria. The expected advantage is an improved ability to take into account the local characteristics of the problem.

A drawback of these methods is that the subproblem (1.2) can usually no longer be solved analytically, since $H_k$ must not be a diagonal matrix. Therefore, part of the investigation of such algorithms is the study of the solutions of these subproblems, usually in combination with a suitable inexactness criterion [25, 62, 63]. Furthermore, the assumption of the Kurdyka-Łojasiewicz property is often used to develop convergence results [29, 31, 32, 64].

**Second-Order Methods** It is only a small step to get from variable metric methods to second-order methods. In these methods, $H_k$ contains second-order information of $f$, i.e. $H_k \approx \nabla^2 f(x^k)$. On the one hand, there exist proximal Newton methods [100, 153, 154, 166, 167], where $H_k = \nabla^2 f(x^k)$, which are known to have the same excellent convergence properties as the Newton method for smooth unconstrained minimization, see e.g. [95]. On the other hand, since this information is computationally expensive for many problems, especially for large-dimensional ones, proximal quasi-Newton or proximal Newton-type methods are considered [15, 16, 68, 81, 82, 91, 95, 114, 119, 145, 153, 154, 169]. Here, $H_k$ is chosen as an appropriate approximation to the Hessian, for example using a quasi-Newton or limited memory quasi-Newton approach. In practice, it is expensive to solve the proximal Newton step exactly, since it is determined by numerical solution of the problem (1.2). Therefore, it is important to understand the convergence of the proximal (quasi-)Newton methods with inexact steps [38, 68, 82, 95, 100, 114, 119, 145, 167].

Another possibility to classify proximal second-order methods is according to the preliminaries required by the particular convergence theory. While in some works [68, 166] only certain concrete problems are investigated, many manuscripts are restricted to the $\ell_1$-norm, i.e $\varphi = \|\cdot\|_1$ [36, 38, 169]. The essential assumption of most methods is convexity of $f$ [15, 16, 68, 95, 145, 153, 154, 169], while some require global Lipschitz continuity of $\nabla f$ [68, 91, 95, 145]. To mitigate these assumptions, error bounds are used [114, 167] or self-concordant functions are considered [100, 153, 154].

Some further remarks on second-order methods are made in the context of the algorithms presented in the subsequent chapters.

**Other Methods** Besides proximal-type methods there are several other approaches and ideas for solving the composite problem (1.1), which should not remain unmentioned. These include semismooth Newton methods [71, 101, 111, 112, 118], trust-region methods [5, 45, 61, 131], interior point methods [87, 88], an alternating direction method of multipliers [97] and fixed point methods [41,43,74,110]. The authors of [26,133,134,148,150] reformulate (1.1) to get a smooth minimization problem with a forward backward envelope, while [69,139] replace the quadratic term in (1.2) by a Bregman distance. If $\varphi$ represents the $\ell_1$-norm, there are orthant wise minimization methods [1, 52, 85], which employ that the $\ell_1$-norm is a linear mapping on every orthant. An overview of several approaches for solving (1.1) is given in [89].

It remains to note that problem (1.1) is also studied in the literature under different assumptions on $f$ and $\varphi$. Worth mentioning are methods where $\varphi$ is nonconvex [34, 35, 64, 70, 72, 83, 165].

## Approaches Presented in this Thesis

In the following chapters, three main approaches to methods for solving (1.1) are investigated. At first, we consider a first-order method in combination with classical line search, which is simply called *proximal gradient method* in the following. The line search is an adaption of the classical Armijo line search, whose purpose is to find $t_k \in (0, 1]$ preferably large such that

$$\psi(x^k + t_k d^k) \leq \psi(x^k) + \sigma t_k \psi'(x^k; d^k)$$

holds for some fixed $\sigma \in (0, 1)$, and update the iterate via $x^{k+1} = x^k + t_k d^k$. We replace the directional derivative by an expression with similar properties, but which is easier to handle. The resulting method is not new, cf. e.g. [156], but known convergence results are collected, edited, and supplemented with additional ones, resulting in a complete survey of the proximal gradient method and its properties.

As mentioned above, first-order methods yield global convergence results under mild assumptions, but a major drawback is that they approximate a solution very slowly, especially for high accuracy. On the other hand, second-order methods have nice local convergence properties, but they are globally convergent only under strongly restrictive (global) assumptions. The main purpose of this thesis is to circumvent both drawbacks by investigating algorithms in the sequel that have both, favorable global and local convergence behavior under mild assumptions. Two methods with different approaches are presented.

The idea of the *globalized inexact proximal Newton-type method* is the combination of an inexact proximal Newton-type method with the proximal gradient method. Whenever possible and a sufficient descent criterion holds, the subproblem (1.2) is solved (inexactly) using a possibly not positive definite matrix $H_k$ containing second-order information on $f$ in $x^k$. Otherwise, $H_k$ is chosen as a positive multiple of the identity matrix and, hence, $d^k$ is obtained by a proximal gradient step. For both possibilities, a line search is performed afterwards.

A different approach is used to obtain the *regularized proximal quasi-Newton method.* Here, the backtracking is based on the idea of ISTA-type methods by, roughly speaking, increasing and reducing $H_k$, if necessary. In detail, we replace the second-order approximation $H_k$ by the regularization $H_k + \mu_k I$ for some $\mu_k > 0$. The regularization parameter $\mu_k$ is adapted in each step by a trust-region approach, depending on the quality of the current step.

We point out that for these methods, an extensive theory for both, global and local convergence, is provided. The global convergence results, which prove convergence to stationary points, need rather mild assumptions. In particular, for the basic results no global Lipschitz continuity of $\nabla f$ is required. Furthermore, the local convergence theory is presented using the Kurdyka-Łojasiewicz property or an error bound assumption, which implies that no strong convexity assumption is needed.

A main difficulty considering variable metric and proximal Newton-type methods is that the solution of the subproblem (1.2) must also be investigated, since this is not possible analytically (in contrast to first-order methods). Therefore, we also address this problem and present an effective method for that purpose, provided $H_k$ is a low rank modification of the identity matrix.

### Structure of the Thesis

The following is an overview of the structure of the thesis. Chapter 2 contains a collection of fundamental results from various fields of analysis, prepared with the purpose of providing a theoretical basis of the following chapters. In particular, this includes concepts of convex and nonsmooth calculus as well as an introduction to the Kurdyka-Łojasiewicz property.

In Chapter 3 we start with introducing the proximity operator and collecting its basic properties. After that, the proximal gradient method is investigated in Section 3.2 and the analysis of global convergence and convergence rates is provided. Afterwards, Section 3.3 is dedicated to the numerical solution of the subproblem (1.2) for special choices of $H_k$.

The following Chapters 4 and 5 deal with two proximal Newton-type methods, namely the globalized inexact proximal Newton-type method in Chapter 4 and the regularized proximal quasi-Newton method in Chapter 5. In both chapters, we start with the deduction of the method, provide an overview of related methods in the literature and state the methods explicitly. After that, the global convergence is investigated under quite mild assumptions and additional results are given for the case that the smooth part $f$ of the objective function has a Lipschitz continuous or uniformly continuous gradient. Finally, local convergence results and convergence rates are deduced. In addition, Chapter 5 closes with some notes on a modified algorithm using a proximal gradient framework in Section 5.5. The basis of these chapters are the research paper [82] and the preprint [81], but substantial effort was undergone to streamline the theory, mitigate the preliminaries, and supplement the investigation by further results.

In Chapter 6 we present extensive numerical material for the methods presented in Chapters 4 and 5, also using the studies of Chapter 3. In particular, the performance of the methods is investigated on common convex and nonconvex problems, and the numerical behaviour of the methods is compared with different state-of-the-art methods.

A final conclusion and some comments on future research topics in Chapter 7 complete the thesis.

# Chapter 2

# Background from Convex and Nonsmooth Analysis

This introductory chapter provides an overview of fundamental concepts and basic results which are essential for the remaining chapters. The majority of the material is a careful collection of results from the literature, selected and arranged to provide a useful overview of the theory needed in the subsequent analysis. Therefore, we skip the proofs of most results and refer to appropriate references.

The following is an outline of the structure of this chapter. In Section 2.1, we start with the basics of convex analysis, in particular the introduction of convexity, convex functions and the properties of the convex subdifferential, which is repeatedly needed for proofs in the following chapters. Section 2.2 covers basic concepts of nonsmooth analysis, in particular, we deal with generalized derivatives, semismooth functions and give a short introduction to the semismooth Newton method. Since they are the main object of this thesis, we collect results for composite functions, which are the sum of a smooth and a convex function, in Section 2.2.3. After that, we introduce the concept of Kurdyka-Łojasiewicz functions in Section 2.3, which can be interpreted as a generalization of strong convexity. Finally, the purpose of Section 2.4 is to introduce some notation related to sequences.

## 2.1 Basics from Convex Analysis

Convexity plays a central role in various areas of mathematics, e.g. optimization, optimal control, calculus of variations and statistics. Even though the concept of convexity is much older, Werner Fenchel [60] was the one who introduced convex analysis as a separate field of mathematics in the 1950s. While his lecture notes remained unpublished for a long time, Rockafellar [141] created one of the outstanding standard works on convex analysis. In the meantime, convex analysis builds a marvellous theoretical framework and plays a fundamental role for modern variational analysis.

We collect the basic concepts of convex functions in Section 2.1.1, summarize existence and uniqueness results for minimizers of (convex) functions in Section 2.1.2 and introduce the convex subdifferential in Section 2.1.3. In addition to the two fundamental publications above, we refer to the monographs of Rockafellar and Wets [142], Bauschke and Combettes [12], and Hiriart-Urruty and Lemaréchal [77] as references for the following results.

### 2.1.1 Convex and Lower Semicontinuous Functions

We start to introduce the concept of convex functions. For that purpose, we recall that a set $C \subset \mathbb{R}^n$ is called *convex* if for any two points in $C$ their connecting line is contained in $C$ or, equivalently, if for all $x, y \in C$ and $\lambda \in (0,1)$, we have $\lambda x + (1-\lambda)y \in C$. This motivates to call a function convex, if its epigraph (roughly speaking the set of all points above its graph) is convex. An equivalent definition is the following.

**Definition 2.1** (Convex function)**.** A function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *convex* on a convex set $C \subseteq \mathbb{R}^n$ if for all $x, y \in C$ and $\lambda \in (0,1)$ there holds

$$\varphi\big(\lambda x + (1-\lambda)y\big) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y). \tag{2.1}$$

Furthermore, the set $\operatorname{dom}(\varphi) := \{x \in \mathbb{R}^n : \varphi(x) < +\infty\}$ is called *domain* of $\varphi$. If $\operatorname{dom}(\varphi) \neq \emptyset$, the function $\varphi$ is called *proper.* The function $\varphi$ is *concave* if $-\varphi$ is convex.

We say that a function is convex, if it is convex on $\mathbb{R}^n$. It is easy to see that the domain of a convex function is convex. Furthermore, a function is convex if and only if it is convex on its domain. Often a stronger version of convexity is needed, which is established next.

**Definition 2.2** (Strong Convexity)**.** A proper function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *strongly convex with modulus* $\mu > 0$ on a convex set $C \subseteq \mathbb{R}^n$ if

$$\varphi\big(\lambda x + (1-\lambda)y\big) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2$$

holds for all $x, y \in C$ and $\lambda \in (0,1)$. The function $\varphi$ is called strongly convex with modulus $\mu > 0$ if $C = \mathbb{R}^n$.

It is trivial to see that strongly convex functions are convex, but the opposite does not hold in general. Strongly convex functions are characterized by the following property.

**Proposition 2.3.** *A proper function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is strongly convex with modulus $\mu > 0$ if and only if $\varphi - \mu\|\cdot\|^2$ is convex.*

We now state the first intriguing property for convex functions.

**Proposition 2.4** (Continuity of convex functions)**.** *A convex function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is continuous relative to any open convex subset of* $\operatorname{dom}(\varphi)$*. In particular, it is continuous relative to* $\operatorname{int}(\operatorname{dom}(\varphi))$*. Any real valued convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is continuous.*

A real valued convex function is therefore Lipschitz continuous on every compact subset of its domain. For the next results we consider differentiable convex functions. Of course, this only makes sense at points for which there exists an entire neighbourhood on which the function is finite valued, i.e. at points in the interior of the domain. For simplicity in the notation, we state these results for real valued convex functions. Since differentiability is a local property, however, the corresponding characterizations only need to hold in an appropriate neighbourhood of the considered points.

**Proposition 2.5** (First-order characterizations)**.** *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then the following hold:*

*(a) $\varphi$ is convex if and only if*

$$\varphi(y) \geq \varphi(x) + \langle \nabla\varphi(x), y - x\rangle \qquad \textit{for all } x, y \in \mathbb{R}^n.$$

*(b) $\varphi$ is strongly convex with modulus $\mu > 0$ if and only if*

$$\varphi(y) \geq \varphi(x) + \langle \nabla\varphi(x), y - x\rangle + \frac{\mu}{2}\|y-x\|^2 \qquad \textit{for all } x, y \in \mathbb{R}^n.$$

A likewise interesting and well-known statement concerns the correlation between convexity and monotonicity of the corresponding gradient mapping.

**Proposition 2.6** (Monotonicity of gradient mappings)**.** *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then the following hold:*

*(a) $\varphi$ is convex if and only if*

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x - y\rangle \geq 0 \qquad \textit{for all } x, y \in \mathbb{R}^n.$$

*(b) $\varphi$ is strongly convex with modulus $\mu > 0$ if and only if*

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x - y\rangle \geq \mu\|x - y\|^2 \qquad \textit{for all } x, y \in \mathbb{R}^n.$$

We now investigate convexity criteria for even twice differentiable functions.

**Proposition 2.7** (Twice differentiable convex functions)**.** *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable. Then the following hold:*

*(a) $\varphi$ is convex if and only if $\nabla^2\varphi(x)$ is positive semidefinite for all $x \in \mathbb{R}^n$.*

*(b) $\varphi$ is strongly convex with modulus $\mu > 0$ if and only if $\nabla^2\varphi(x) \succeq \mu I$ holds for all $x \in \mathbb{R}^n$.*

Although the previous results seem useful, continuity, and therefore differentiability, does not need to hold outside the interior of the domain. As a consequence, the theory of convex functions is most powerful in the presence of lower semicontinuity, which is introduced next.

**Definition 2.8** (Lower semicontinuity)**.** A function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *lower semicontinuous* at $x \in \mathbb{R}^n$ if every sequence $\{x^k\}$ converging to $x$ satisfies

$$\liminf_{k\to\infty} \varphi(x^k) \geq \varphi(x).$$

The function $\varphi$ is called *lower semicontinuous* if it is lower semicontinuous at every point in $\mathbb{R}^n$.

A first consequence of the concept of lower semicontinuity is the following, which explains why lower semicontinuous functions are also called closed by some authors.

**Proposition 2.9.** *A function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is lower semicontinuous if and only if for all $\alpha \in \mathbb{R}$ the level sets $\operatorname{lev}_{\leq\alpha} \varphi$ are closed.*

Many elementary operations preserve convexity and lower semicontinuity. The most basic ones are stated in the following result.

**Proposition 2.10.** *Let $\varphi_1, \varphi_2 : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex, proper and lower semicontinuous functions and $A : \mathbb{R}^m \to \mathbb{R}^n$ a linear mapping. Then the sum $\varphi_1 + \varphi_2$ and the composition $\varphi_1 \circ A$ are convex and lower semicontinuous. They are proper, if $\operatorname{dom}(\varphi_1) \cap \operatorname{dom}(\varphi_2) \neq \emptyset$ or $\operatorname{dom}(\varphi_1) \cap \operatorname{Im}(A) \neq \emptyset$, respectively.*

We are mainly interested in convex, proper and lower semicontinuous functions, since the set of these functions is the one dealt with in problem (1.1). Hence, for the theory in the subsequent chapters we consider functions which have all three of these properties. It is worth noting, however, that some of the results hold without assuming lower semicontinuity. At this point we refer the reader to the above mentioned standard literature of convex analysis.

### 2.1.2 Minimization and Convexity

In this thesis, convexity is used in particular in the context of optimization problems. To address that purpose we define coercivity and give basic results on minimizers of coercive and convex functions.
We start with the fact that for convex functions a separation of local and global minima is not necessary. For this, recall that a global minimizer of $\varphi$ is a point with optimal function value, while a local minimizer minimizes the function only in some neighbourhood. The next result is about the set of all minimizers of a convex function $\varphi$, which is denoted by $\arg\min \varphi$.

**Proposition 2.11.** *Every local minimizer of a proper and convex function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a global minimizer.*

This result highlights the fundamental interest and importance of convexity for minimization problems. In particular, in combination with the following one, the result ensures that descent methods, which are proven to find a local minimum, always find the global minimum, independently of the initialization of the method. This is a major benefit of considering optimization problems with convex functions.

**Proposition 2.12.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex. Then the set $\arg\min \varphi$ of (global) minimizers of $\varphi$ is convex.*

In general, the set of minimizers of a convex function may contain more than one point. However, the following result shows that this is not possible for strongly convex functions.

**Proposition 2.13.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and strongly convex. Then $\varphi$ has at most one minimizer.*

So far we have considered the set of minima of a (strongly) convex function. However, this set might be empty. For that reason we introduce the concept of coercivity, under which convex functions always have minima.

**Definition 2.14.** A function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *coercive* if

$$\lim_{\|x\| \to \infty} \varphi(x) = +\infty,$$

and *supercoercive* if

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

Note that the above properties are not used consistently in the literature. Instead of coercive and supercoercive, some authors use the terms 0-coercive and 1-coercive. We continue with a characterization of convex coercive functions.

**Proposition 2.15.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex. Then the following are equivalent:*
*(a) $\varphi$ is coercive,*
*(b) for every $\alpha \in \mathbb{R}$ the level set $\operatorname{lev}_{\leq\alpha} \varphi$ is bounded,*
*(b) there exists $\alpha \in \mathbb{R}$ such that the level set $\operatorname{lev}_{\leq\alpha} \varphi$ is nonempty and bounded.*

Note that the equivalence of (a) and (b) even holds without convexity, while the characterization in (c) is based on the assumption that $\varphi$ is convex. The following two results show the existence of minimizers of convex functions. With these we conclude the section.

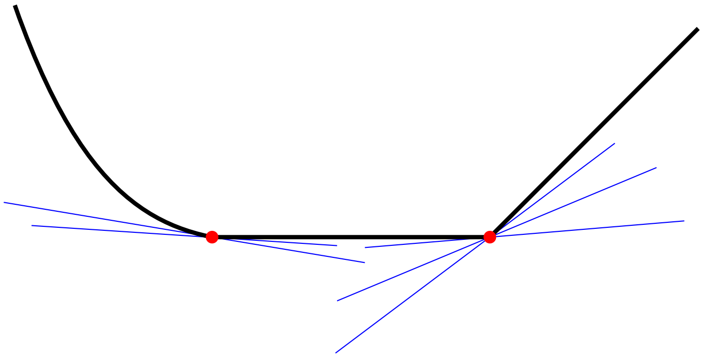


Figure 2.1: Example of a convex function and some of its subgradients, represented as affine minorants

**Proposition 2.16.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex, proper and lower semicontinuous and let $C \subseteq \mathbb{R}^n$ be a closed convex set such that $C \cap \operatorname{dom} \varphi \neq \emptyset$. If $\varphi$ is coercive, it has a minimizer over $C$.*

**Corollary 2.17.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be strongly convex, proper and lower semicontinuous. Then $\varphi$ is supercoercive and has exactly one minimizer.*

### 2.1.3 The Convex Subdifferential

Even though they are not differentiable in general, convex functions have many useful differentiability properties. The central concept of the convex subdifferential, which will be introduced in the following, is a generalization of the gradient of a smooth function and a fundamental tool in the analysis of nondifferentiable convex functions. If a convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is differentiable, Proposition 2.5(a) states that

$$\varphi(y) \geq \varphi(x) + \langle s, y - x \rangle$$

holds for all $x, y \in \mathbb{R}^n$ with $s := \nabla\varphi(x)$. Geometrically, this means that the tangential hyperplane $\varphi(x) + \langle s, y - x \rangle$, which coincides with $\varphi$ in $x$, minorizes $\varphi$. For nondifferentiable convex functions, a subgradient is defined by one such tangential hyperplane, see Figure 2.1, and the set of all tangential hyperplanes in one point leads to the convex subdifferential.

**Definition 2.18** (Convex Subdifferential)**.** Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex and $x \in \mathbb{R}^n$. Then, $s \in \mathbb{R}^n$ is called *subgradient* of $\varphi$ at $x$, if

$$\varphi(y) \geq \varphi(x) + \langle s, y - x \rangle$$

holds for all $y \in \mathbb{R}^n$. The set of all subgradients of $\varphi$ at $x$ is the *(convex) subdifferential* $\partial\varphi(x)$, and we say that $\varphi$ is *subdifferentiable* at $x$ if $\partial\varphi(x) \neq \emptyset$.

We note that the concept of subgradients and the subdifferential can in principle be transferred to nonconvex functions with the above definition. However, it is possible that in some points the function is not subdifferentiable in that case, while the following result holds for convex functions.

**Proposition 2.19.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex, and let $x \in \mathbb{R}^n$. Then the following hold:*

*(a) $\partial\varphi(x) \neq \emptyset$, if $x \in \operatorname{int}(\operatorname{dom}\varphi)$,*

*(b) $\partial\varphi(x) = \emptyset$, if $x \notin \operatorname{dom}\varphi$,*

*(c) $\partial\varphi(x)$ is a closed, convex set.*

Furthermore, if $x$ is an element of the boundary of $\operatorname{dom}\varphi$ and $\operatorname{int}(\operatorname{dom}\varphi)$ is nonempty, it can be shown that $\partial\varphi(x)$ is either empty or unbounded. We continue with some basic examples for subdifferentials of convex functions, which can easily be verified.

**Example 2.20.**

(a) (Euclidean norm) For the Euclidean norm $\|\cdot\|$ we have

$$\partial\|\cdot\|(x) = \begin{cases} \frac{x}{\|x\|} & \text{if } x \neq 0, \\ \overline{B}_1(0) & \text{if } x = 0. \end{cases}$$

(b) ($\ell_1$-norm) The subdifferential of the $\ell_1$-norm is given element-wise by

$$\big(\partial\|\cdot\|_1(x)\big)_i = \begin{cases} 1 & \text{if } x_i > 0, \\ -1 & \text{if } x_i < 0, \\ [-1,1] & \text{if } x_i = 0 \end{cases}$$

for $i = 1, \ldots, n$.

(c) (Indicator function) Let $C \subset \mathbb{R}^n$ be a closed convex set and $x \in C$. Then the definition of subgradients yields $s \in \partial\iota_C(x)$ if and only if $\iota_C(y) \geq \iota_C(x) + \langle s, y - x\rangle$ holds for all $y \in \mathbb{R}^n$, which is equivalent to $0 \geq \langle s, y - x\rangle$ for all $y \in C$. Hence,

$$\partial\iota_C(x) = N_C(x) := \big\{s \in \mathbb{R}^n \mid \langle s, y - x\rangle \leq 0 \text{ for all } y \in C\big\}$$

is the so called *normal cone* of $C$ in $x$. For $x \notin C$, Proposition 2.19 yields $\partial\iota_C(x) = \emptyset$.

(d) (Half circle) Consider the function $\varphi : \mathbb{R} \to \overline{\mathbb{R}}$,

$$\varphi(x) = \begin{cases} -\sqrt{1 - x^2} & \text{if } |x| \leq 1, \\ +\infty & \text{otherwise}, \end{cases}$$

which describes a lower half circle. Then $\partial\varphi(x) = \{\varphi'(x)\}$ for $|x| < 1$ and $\partial\varphi(x) = \emptyset$ for $|x| \geq 1$. In particular, $\pm 1 \in \operatorname{dom}\varphi$ and $\partial\varphi(\pm 1) = \emptyset$. ◊

To determine the subdifferential of various convex functions that come out of convexity-preserving operations, there are several calculus rules. We collect some elementary ones in the following result.

**Proposition 2.21** (Subdifferential calculus)**.**

*(a) (Separable functions) Assume that $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, convex and separable, i.e. there exist proper, convex functions $\varphi_i : \mathbb{R} \to \overline{\mathbb{R}}$ for $i = 1, \ldots, n$ such that $\varphi(x) = \sum_{i=1}^n \varphi_i(x_i)$. Then*

$$\partial\varphi(x) = \partial\varphi_1(x_1) \times \cdots \times \partial\varphi_n(x_n) \qquad \text{for all } x \in \mathbb{R}^n.$$

*(b) (Multiples) Let $\lambda > 0$ and $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex. Then*

$$\partial(\lambda\varphi_1)(x) = \lambda\partial\varphi_1(x) \qquad \text{for all } x \in \mathbb{R}^n.$$

*(c) (Translation) For* $b \in \mathbb{R}^n$ *and a proper, convex function* $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *let* $\overline{\varphi}(x) := \varphi(x-b)$*. Then*

$$\partial\overline{\varphi}(x) = \partial\varphi(x-b) \qquad \textit{for all } x \in \mathbb{R}^n.$$

*(d) (Sum rule) Let* $\varphi_1, \varphi_2 : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be proper and convex. If* $\operatorname{int}(\operatorname{dom}\varphi_1) \cap \operatorname{int}(\operatorname{dom}\varphi_2) \neq \emptyset$*, there holds*

$$\partial(\varphi_1 + \varphi_2)(x) = \partial\varphi_1(x) + \partial\varphi_2(x) \qquad \textit{for all } x \in \mathbb{R}^n.$$

*(e) (Chain rule) Let* $A \in \mathbb{R}^{m\times n}$*,* $\varphi : \mathbb{R}^m \to \overline{\mathbb{R}}$ *be a proper, convex function and suppose* $\operatorname{Im}(A) \cap \operatorname{int}(\operatorname{dom}\varphi_1) \neq \emptyset$*. Then*

$$\partial(\varphi_1 \circ A)(x) = A^T(\partial\varphi_1 \circ A)(x) \qquad \textit{for all } x \in \mathbb{R}^n.$$

Note that Proposition 2.21(a) justifies the computation of the subdifferential of the $\ell_1$-norm in Example 2.20(b), which is given element-wise.
We deduced the idea of subgradients from Proposition 2.5(a). Inspired from part (b) of that result, we obtain the following characterizations of strongly convex functions.

**Proposition 2.22.** *The following statements are equivalent for a proper, convex function* $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *and some* $\mu > 0$*:*

*(a)* $\varphi$ *is strongly convex with modulus* $\mu$*.*

*(b) The inequality*

$$\varphi(y) \geq \varphi(x) + \langle s, y-x\rangle + \frac{\mu}{2}\|x-y\|^2$$

*holds for all* $x, y \in \mathbb{R}^n$ *and* $s \in \partial\varphi(x)$*.*

*a) The subdifferential of* $\varphi$ *is strongly monotone with modulus* $\mu$*, i.e.*

$$\langle s - \tilde{s}, x-y\rangle \geq \mu\|x-y\|^2$$

*holds for all* $x, y \in \mathbb{R}^n$ *and* $s \in \partial\varphi(x)$*,* $\tilde{s} \in \partial\varphi(y)$*.*

Two more substantial properties of proper, convex and lower semicontinuous functions are stated in the following, since they are essential for several proofs involving the convex subdifferential: The outer semicontinuity or closedness of the subdifferential, and the property that it maps bounded sets to bounded sets.

**Proposition 2.23** (Outer Semicontinuity)**.** *Let* $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be proper, convex and lower semicontinuous. If* $\{x^k\} \subset \mathbb{R}^n$ *and* $\{s^k\} \subset \mathbb{R}^n$ *are sequences such that* $s^k \in \partial\varphi(x^k)$ *for all* $k \in \mathbb{N}$*, where* $\{x^k\}$ *converges to* $x^*$ *and* $\{s^k\}$ *converges to* $s^*$*, then* $s^* \in \partial\varphi(x^*)$*.*

**Proposition 2.24** (Boundedness)**.** *Let* $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a proper, convex and lower semicontinuous function, and let* $C \subset \operatorname{int}(\operatorname{dom}\varphi)$ *be nonempty, closed and bounded. Then the set*

$$\partial\varphi(C) := \bigcup\big\{\partial\varphi(x) \mid x \in C\big\}$$

*is nonempty, closed and bounded. Furthermore,* $\varphi$ *is Lipschitz continuous with Lipschitz constant*

$$L := \sup\big\{\|s\| \mid s \in \partial\varphi(C)\big\} < +\infty$$

*on the set* $C$*.*

A very elementary, but essential result on global minimizers of proper convex functions is their characterization by the subdifferential. The origin of this powerful idea traces back to the work of Pierre de Fermat in the 17th century, and therefore it is known as *Fermat's* *rule* in convex optimization. We note that the same result also holds without convexity, but, to be consistent with Definition 2.18, the result is formulated under the assumption of convexity.

**Proposition 2.25** (Fermat's rule). *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex. Then*

$$\arg\min \varphi = \{x \in \mathbb{R}^n \mid 0 \in \partial\varphi(x)\}.$$

For the remainder of this section, we study the connection between convexity and differentiability. It turns out that convex functions have a number of useful differentiability properties connected to the subdifferential. If a convex function is differentiable, its gradient is a subgradient, see the deduction at the beginning of this section. The following result generalizes this observation.

**Proposition 2.26.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, convex function, and let $x \in \operatorname{dom}\varphi$. If $\varphi$ is differentiable at $x$, then $\nabla f(x)$ is the unique subgradient of $\varphi$ at $x$. Conversely, if $\varphi$ has a unique subgradient at $x$, then $\varphi$ is differentiable at $x$.*

A useful property of convex functions is the fact that one-sided directional derivatives exist universally. To formally state this result, we first define the directional derivative.

**Definition 2.27.** Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper. For $x \in \operatorname{dom}\varphi$ we say that $\varphi$ is *directionally differentiable* at $x$ in direction $d \in \mathbb{R}^n$ if the limit

$$\lim_{t \downarrow 0} \frac{\varphi(x + td) - \varphi(x)}{t}$$

exists. In this case we call

$$\varphi'(x; d) := \lim_{t \downarrow 0} \frac{\varphi(x + td) - \varphi(x)}{t}$$

the *directional derivative* of $\varphi$ at $x$ in the direction of $d$.

With this definition it is possible to summarize some properties of the difference quotient and the directional derivative of a convex function.

**Proposition 2.28.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex, $x \in \operatorname{dom}\varphi$ and $d \in \mathbb{R}^n$. Then the following hold:*

*(a) The difference quotient*

$$t \mapsto \frac{\varphi(x + td) - \varphi(x)}{t}$$

*is nondecreasing for $t > 0$.*

*(b) The directional derivative $\varphi'(x; d)$ exists (in $\overline{\mathbb{R}}$) with*

$$\varphi'(x; d) = \inf_{t > 0} \frac{\varphi(x + td) - \varphi(x)}{t}.$$

*(c) $\varphi'(x; \cdot)$ is sublinear (and therefore convex) and proper.*

*(d) $\varphi'(x; \cdot)$ is lower semicontinuous for $x \in \operatorname{int}(\operatorname{dom}\varphi)$.*

**Remark 2.29.** Considering a continuously differentiable function $\varphi$ for the moment, we know that the directional derivative $\varphi'(x; d)$ coincides with the expression $\nabla\varphi(x)^T d$. Hence, the directional derivative $\varphi'$ is continuous as a function of $(x, d)$. In contrast, this does not hold for nondifferentiable convex functions. As an example, consider the one-dimensional mapping $\varphi(x) = |x|$. Then,

$$\varphi'(x;d) = \begin{cases} d, & \text{if } x > 0, \\ |d|, & \text{if } x = 0, \\ -d, & \text{if } x < 0, \end{cases}$$

is not continuous in $x = 0$. Investigating the convergence of optimization methods for convex functions, this issue causes some difficulties, which are addressed in the subsequent chapters. ◊

We close this section with a generalization of the identity $\varphi'(x;d) = \nabla\varphi(x)^T d$ to nondifferentiable functions $\varphi$, which yields a connection between the directional derivative and the convex subdifferential.

**Proposition 2.30.** *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and convex and $x \in \operatorname{dom}\varphi$. Then the following are equivalent:*

$$s \in \partial\varphi(x) \qquad \Longleftrightarrow \qquad \varphi'(x;d) \geq \langle s, d\rangle \quad \text{for all } d \in \mathbb{R}^n.$$

*If $x \in \operatorname{int}(\operatorname{dom}\varphi)$, we have in addition*

$$\varphi'(x;\cdot) = \max_{s\in\partial\varphi(x)} \langle s, \cdot\rangle.$$

## 2.2 Basics from Nonsmooth Analysis

Nonsmooth calculus is a collective term for a wide field of calculus whose common purpose is the study of nondifferentiable functions. The class of problems we consider in (1.1) is among these functions. We therefore address some basics of nonsmooth analysis in this section, where we concentrate the discussion to locally Lipschitz continuous functions. Recall that a function $F : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz continuous near $x \in \mathbb{R}^n$ if there exist $\varepsilon > 0$ and a Lipschitz constant $L > 0$ (depending on $x$) such that

$$\|F(y_1) - F(y_2)\| \leq L\|y_1 - y_2\| \qquad \text{for all } y_1, y_2 \in B_\varepsilon(x).$$

This analysis includes the definition of generalized Jacobians and their properties in Section 2.2.1, as well as the concepts of Newton derivatives and semismooth functions in Section 2.2.2. At the end of this section, we briefly discuss the semismooth Newton method, before we focus on some results regarding composite functions in the form of (1.1) in Section 2.2.3. Some fundamental references for nonsmooth analysis in literature are the monograph by Clarke [47] and the book of Ulbrich [159]. In addition to these references, we mention [48], in particular for Newton derivatives [44], and [136] for the semismooth Newton method.

### 2.2.1 Generalized Nonsmooth Subdifferentials

The purpose of this section is to introduce the generalized Jacobian by Clarke, which covers both, the derivatives of smooth functions, and the subdifferentials of convex functions introduced in Section 2.1.3. Since derivatives require some kind of smoothness assumption, the following results are stated for locally Lipschitz continuous functions. According to Rademacher's theorem (see e.g. [75, Theorem 3.1]), such functions are almost everywhere differentiable, which means that the set of points, where the function is not differentiable, has Lebesgue measure zero. Hence, the following makes sense.

**Definition 2.31** (Generalized Subdifferentials)**.** Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous in a neighbourhood of $x \in \mathbb{R}^n$. The set

$$\partial_B F(x) := \{M \in \mathbb{R}^{m\times n} \mid \exists\{x^k\} \subset \mathcal{D}_F : x^k \to x, F'(x^k) \to M\}$$

is called *Bouligand-* or *B-subdifferential* of $F$ at $x$, where $\mathcal{D}_F \subset \mathbb{R}^n$ denotes the set of all points, where $F$ is differentiable. Moreover, the convex hull $\partial F(x) := \operatorname{conv}(\partial_B F(x))$ denotes the *generalized Jacobian in the sense of Clarke* of $F$ at $x$.

If $m = 1$, the generalized Jacobian $\partial F(x)$ is referred to as *generalized gradient* as well. Let us illustrate the definition with some examples.

**Example 2.32.**

(a) (Euclidean norm) Let $F(x) = \|x\|_2$. Then we get $\partial F(x) = \partial_B F(x) = \{\nabla F(x)\}$ for all $x \neq 0$. Let $x = 0$ and note that for $y \neq 0$ we have $\nabla F(y) = y/\|y\|_2$. Hence, $\partial_B F(0) \subseteq \{s \mid \|s\|_2 = 1\}$. On the other hand, let $s \in \mathbb{R}^n$ with $\|s\|_2 = 1$ be given. Then, the sequence $\{\frac{1}{k}s\}_{k\in\mathbb{N}}$ converges to 0 and the gradient of $F$ on that sequence is constantly equal to $s$. Hence, we get

$$\partial_B F(0) = \{s \mid \|s\|_2 = 1\} \qquad \text{and} \qquad \partial F(0) = \{s \mid \|s\|_2 \leq 1\}.$$

(b) ($\ell_1$-norm) Let $F(x) = \|x\|_1$. It is easy to see that the B-subdifferential can be computed element-wise for this example (but not in general). Hence, for $i = 1, \dots, n$ we get

$$\partial_B F(x)_i = \begin{cases} \{1\}, & \text{if } x_i > 0, \\ \{-1\}, & \text{if } x_i < 0, \\ \{-1, 1\}, & \text{if } x_i = 0, \end{cases} \qquad \text{and} \qquad \partial F(x)_i = \begin{cases} \{1\}, & \text{if } x_i > 0, \\ \{-1\}, & \text{if } x_i < 0, \\ [-1, 1], & \text{if } x_i = 0. \end{cases}$$

(c) (Differentiable function) Let $F(x) = x^2 \sin(1/x)$. The function is differentiable with

$$F'(x) = \begin{cases} 2x\sin(1/x) - \cos(1/x), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0. \end{cases}$$

In particular, this yields $0 \in \partial_B F(0)$ by choosing a sequence, which is constantly 0. Using the vanishing sequences $\{1/(2\pi k + \alpha)\}$ for $\alpha \in [0, 2\pi]$ and $k \geq 0$ yields $\partial_B F(0) = \partial F(0) = [-1, 1]$. ◊

The differentials $\partial_B F$ and $\partial F$ have, inter alia, the following properties.

**Proposition 2.33.** *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous and $x \in \mathbb{R}^n$. Then the following hold:*

*(a) $\partial_B F(x)$ is nonempty and compact.*

*(b) $\partial F(x)$ is nonempty, compact and convex.*

*(c) $\partial_B F(x) \subseteq \partial F(x)$.*

*(d) $\partial F$ is closed, i.e. for sequences $\{x^k\} \subset \mathbb{R}^n$ converging to $x^*$ and $M_k \in \partial F(x^k)$ converging to $M^* \in \mathbb{R}^{m\times n}$ there holds $M^* \in \partial F(x^*)$.*

Furthermore, as mentioned above, these are indeed generalizations of the derivatives of smooth functions and the subdifferentials of convex functions. Hence, there is no clash of notation for the convex subdifferential and Clarke's generalized Jacobians.

**Proposition 2.34.**

*(a) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable in a neighbourhood of $x \in \mathbb{R}^n$. Then*

$$\partial F(x) = \partial_B F(x) = \{F'(x)\}.$$

*(b) Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous and $x \in \operatorname{int}(\operatorname{dom} \varphi)$. Then Clarke's generalized Jacobian coincides with the subdifferential at $x$ in the sense of convex analysis (Definition 2.18).*

We note that in Proposition 2.34(b), and from now on, we are slightly inconsistent in notation: If a function $F : \mathbb{R}^n \to \mathbb{R}$ is given, the differentials in Definition 2.31 are subsets of $\mathbb{R}^{1\times n}$ (row vectors) but the elements of the convex subdifferential are considered as column vectors, which is the usual convention for functions mapping to the real numbers. This distinction is irrelevant as we remain in the case $m = 1$, but must be adhered to in interpreting, so for the following chain rule.

**Proposition 2.35** (Chain rule for generalized Jacobians)**.** *Let $h = g \circ F$, where $F : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous near $x \in \mathbb{R}^n$ and $g : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous near $F(x)$. Then $h$ is Lipschitz continuous in a neighbourhood of $x$ and one has*

$$\partial h(x) \subseteq \operatorname{conv}\big(\partial g(F(x))\partial F(x)\big).$$

*If, in addition, $g$ is continuously differentiable at $F(x)$, then equality holds, and the convex hull is superfluous.*

We note that several more calculus rules for $\partial F$ can be found in the literature, while there are hardly any rules for the B-subdifferential $\partial_B F$. Since these are not necessary for the following chapters, we skip stating them explicitly. Similar to Proposition 2.25 we conclude the section with Fermat's principle, this time in the version for nonsmooth functions.

**Proposition 2.36** (Fermat's rule for nonsmooth functions)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz continuous in a neighbourhood of $x \in \mathbb{R}^n$. If $x$ is a local minimizer of $F$, then $0 \in \partial F(x)$.*

Note that there is no Fermat's principle for the B-subdifferential. Furthermore, in contrast to convex analysis, $0 \in \partial F(x)$ is no sufficient condition for local or global minimizers. In particular, the same inclusion holds for a local maximum.

### 2.2.2 Newton Derivatives and the Semismooth Newton Method

The notion of semismoothness in the context of functions $F : \mathbb{R}^n \to \mathbb{R}^m$ was introduced and investigated by Qi [135] and Qi and Sun [136]. Later on the definition was extended to so called Newton differentiable or slantly differentiable functions, see [44]. The importance of Newton derivatives comes from the fact that, although the underlying mapping is in general nonsmooth, Newton's method is still applicable and converges locally at a superlinear rate. In the following, we give a brief introduction to these terms and their relations, and state the semismooth Newton method explicitly.

**Definition 2.37** (Newton differentiability)**.** A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is called *Newton (or slantly) differentiable* at $x \in \mathbb{R}^n$ if there exists a mapping $D_{N,x}F : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ such that

$$\lim_{s\to 0} \frac{\|F(x+s) - F(x) - D_{N,x}F(x+s)s\|}{\|s\|} = 0.$$

The mapping $D_{N,x}$ is called a *Newton derivative or slanting function* for $F$ in $x$. We say that $F$ is *Newton differentiable* if it is Newton differentiable in every $x \in \mathbb{R}^n$.

We refer to [44] for a detailed discussion and several comments on the properties of Newton differentiable functions. For our purpose, it is sufficient to highlight just a few traits. In particular, we point out that Newton derivatives are not unique. Furthermore, there must not exist one single mapping $G : \mathbb{R}^n \to \mathbb{R}^{m\times n}$, which is a Newton derivative for all $x \in \mathbb{R}^n$, but the mapping depends on $x$. It follows immediately from the definition that a continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}^m$ is Newton differentiable with Newton derivative $D_{N,x}F = F'$ for all $x \in \mathbb{R}^n$, but there are several more candidates. For example, the constant mapping $D_{N,x}F : x + s \mapsto F'(x)$ is a Newton derivative for $F$ in $x$. An important result is [44, Theorem 2.6], which shows that lots of functions are Newton differentiable.

**Proposition 2.38.** *A mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is Newton differentiable if and only if $F$ is locally Lipschitz continuous.*

Although this result is very basic and the proof is even constructive, it is not very useful numerically, since in addition to the Newton derivative itself, for the application of a nonsmooth Newton method, see below, its inverse is needed. Therefore, the corresponding matrix should have a special structure and be easy to invert, and a Newton derivative should be chosen to satisfy this property. In some cases, this is achieved using semismooth functions, which provide a connection to the nonsmooth subdifferentials introduced in the previous section. For that purpose, we briefly introduce semismoothness.

**Definition 2.39** (Semismooth function)**.** A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is called *semismooth* at $x \in \mathbb{R}^n$ if it is Lipschitz continuous in a neighbourhood of $x$ and the limit

$$\lim_{\substack{H\in\partial F(x+\tau d)\\ d\to s,\, \tau\downarrow 0}} Hd$$

exists for all $s \in \mathbb{R}^n$. If $F$ is semismooth at all $x \in \mathbb{R}^n$, we call $F$ *semismooth*.

The local Lipschitz continuity is included in this definition. Hence, if a function is semismooth in $x$, it is also Lipschitz continuous near $x$. Semismoothness admits some characterizations. One of them is the following.

**Proposition 2.40.** *A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is semismooth at $x \in \mathbb{R}^n$ if and only if it is Lipschitz continuous near $x$, the directional derivative $F'(x;\cdot)$ exists, and*

$$\limsup_{\substack{s\to 0\\ H\in\partial F(x+s)}} \frac{\|F(x+s) - F(x) - Hs\|}{\|s\|} = 0.$$

This result shows once more that every semismooth function is Newton differentiable. If $F : \mathbb{R}^n \to \mathbb{R}^m$ is semismooth in $x$, every mapping $D_{N,x}F : y \mapsto H(y)$, where $H(y) \in \partial F(y)$, is a Newton derivative of $F$ in $x$. A similar result holds for elements of the B-subdifferential $\partial_B F(x)$ since $\partial_B F(x) \subseteq \partial F(x)$.

It is easy to see that Newton derivatives are linear, i.e. if $F, G : \mathbb{R}^n \to \mathbb{R}^m$ are Newton differentiable in $x \in \mathbb{R}^n$ with Newton derivatives $D_{N,x}F, D_{N,x}G$ as well as $\alpha, \beta \in \mathbb{R}$, then $\alpha F + \beta G$ is Newton differentiable in $x$ with Newton derivative $\alpha D_{N,x}F + \beta D_{N,x}G$. This makes it easy to calculate Newton derivatives of composite functions. On the other hand, an analogous result in this generality does not hold for semismooth functions based on Clarke's generalized Jacobian [47].

Besides the linearity, further tools for the calculation of Newton derivatives are two more chain rules, where we note that the second one is actually a special case of the first one without the assumption of uniform boundedness.

**Proposition 2.41.**

*(a) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be Newton differentiable at $x \in \mathbb{R}^n$ with Newton derivative $D_{N,x}F$ and $G : \mathbb{R}^m \to \mathbb{R}^p$ be Newton differentiable at $y = F(x)$ with Newton derivative $D_{N,y}G$. If $D_{N,x}F$ and $D_{N,y}G$ are uniformly bounded in a neighbourhood of $x$ and $y$, respectively, then $G \circ F$ is also Newton differentiable at $x$ with Newton derivative*

$$D_{N,x}(G \circ F) = D_{N,y}G \circ D_{N,x}F.$$

*(b) Let $F : \mathbb{R}^m \to \mathbb{R}^p$ be Newton differentiable in $Ax + y$ with Newton derivative $D_{N,Ax+y}F$, where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$. Then $G(u) := F(Au+y)$ is Newton differentiable in $x$ with Newton derivative $D_{N,x}F \circ A$.*

With this on hand, we consider the semismooth Newton method, first introduced by Qi and Sun [136], in its most basic form. Suppose that we are looking for a zero of a Newton differentiable function $F : \mathbb{R}^n \to \mathbb{R}^n$, i.e. we want to find $x^* \in \mathbb{R}^n$ such that $F(x^*) = 0$. The semismooth Newton method (in analogy to the smooth Newton method) to solve this problem is stated in Algorithm 2.1.

**Algorithm 2.1** Semismooth Newton Method

```
(S.0) Choose $x^0 \in \mathbb{R}^n$, and set $k := 0$.
(S.1) If $F(x^k) = 0$: STOP.
(S.2) Choose $H_k \in \mathbb{R}^{n\times n}$.
(S.3) Compute the Newton step $d^k$ as solution of $H_k d^k = -F(x^k)$.
(S.4) Set $x^{k+1} := x^k + d^k$, $k \leftarrow k + 1$, and go to (S.1).
```

Note that the matrix $H_k$ in (S.2) is ideally chosen to be $D_{N,x^*}F(x^k)$, where $D_{N,x^*}F$ is a Newton derivative of $F$ at $x^*$. Unless $x^*$ is already known, an appropriate approximation is needed. The sufficient condition for this approximation is stated in the following convergence result, cf. [44, Theorem 3.4], which is similar to the one for the classical Newton method.

**Proposition 2.42.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be Newton differentiable at $x^* \in \mathbb{R}^n$ with $F(x^*) = 0$. Moreover, let $D_{N,x^*}F$ be a Newton derivative of $F$ at $x^*$ and $\|D_{N,x^*}F(x)^{-1}\| \leq M$ in a neighbourhood of $x^*$ for some $M > 0$. Then the sequence $\{x^k\}$ generated by Algorithm 2.1 converges superlinearly to $x^*$ in a neighbourhood of $x^*$, i.e.*

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0,$$

*if the matrices $H_k = H(x^k)$ satisfy $\|H(x^* + s) - D_{N,x^*}F(x^* + s)\| \to 0$ for $\|s\| \to 0$.*

### 2.2.3 Some Results on Composite Functions

Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, convex and lower semicontinuous function and assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable in an open set containing $\operatorname{dom}\varphi$. Set $\psi := f + \varphi$. We study the construction of the subdifferential of such composite functions in this section, since they form the integral part of problem (1.1). Since $f$ is not necessarily convex and $\varphi$ might not be finite everywhere, neither the results for the convex subdifferential (Section 2.1.3) nor for the generalized gradient in the sense of Clarke (Section 2.2.1) apply directly in this setting. Thus, we have to investigate this topic separately. Although also studied by Clarke [47, Chapter 2], we follow the approach of Rockafellar and Wets [142, Chapter 8].

Note that a short computation and Proposition 2.28 (b) show that for any $x \in \operatorname{dom}\varphi$ and $d \in \mathbb{R}^n$ the directional derivative $\psi'(x;d)$ exists, and we get the identity

$$\psi'(x;d) = f'(x;d) + \varphi'(x;d) = \nabla f(x)^T d + \varphi'(x;d). \tag{2.2}$$

Hence, the following definition of stationary points is natural.

**Definition 2.43** (Stationary point)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. We call $x \in \operatorname{dom}\varphi$ a *stationary point* of $\psi = f + \varphi$ if for all directions $d \in \mathbb{R}^n$ the directional derivative $\psi'(x;d)$ is nonnegative, i.e. $\psi'(x;d) \geq 0$.

In the subsequent chapters we investigate the convergence of the stated algorithms to stationary points. A fundamental tool for that is Fermat's rule adjusted to the preliminaries. An important step for the proof of this result is the definition of subgradients of $\psi$.

**Definition 2.44.** Consider a function $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $x \in \operatorname{dom}\psi$. We say that $s \in \mathbb{R}^n$ is

(a) a *regular subgradient* of $\psi$ at $x$, written $s \in \hat{\partial}\psi(x)$, if

$$\psi(y) \geq \psi(x) + \langle s, y - x\rangle + o(\|y - x\|) \quad \text{for } y \to x,$$

(b) a *(general) subgradient* of $\psi$ at $x$, written $s \in \partial\psi(x)$, if there are sequences $\{x^k\}$ converging to $x$ and $\{s^k\}$ converging to $s$ such that $s^k \in \hat{\partial}\psi(x^k)$ holds for all $k \in \mathbb{N}$.

For convex functions, these definitions of subgradients are further characterizations of the convex subdifferential, as shown in the ensuing result. Hence, again, there is no clash in notation.

**Proposition 2.45.** *For any proper, convex function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ and any point $x \in \operatorname{dom}\varphi$, one has*

$$\hat{\partial}\varphi(x) = \partial\varphi(x) = \{s \in \mathbb{R}^n \mid \varphi(y) \geq \varphi(x) + \langle s, y - x\rangle \textit{ for all } y \in \mathbb{R}^n\}.$$

Combining this fact with the definition of the general subgradients yields the sum rule for the subdifferential of composite functions.

**Lemma 2.46.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous, and set $\psi = f + \varphi$. For $x \in \mathbb{R}^n$ there holds*

$$\partial\psi(x) = \nabla f(x) + \partial\varphi(x) := \big\{\nabla f(x) + s \mid s \in \partial\varphi(x)\big\}.$$

With this crucial identity in hand, a characterization of stationary points follows directly from their definition, which is again in the form of Fermat's rule.

**Proposition 2.47** (Fermat's rule for composite functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. A point $x \in \mathbb{R}^n$ is a stationary point of $\psi = f + \varphi$ if and only if $0 \in \nabla f(x) + \partial\varphi(x)$.*

One more useful property of the studied composite functions concerns their directional derivatives. In addition to the considered limit in Definition 2.27, the following more general statement holds, which result from the combination of [47, Proposition 2.3.6] and some elementary calculations.

**Proposition 2.48.** *Let $x^* \in \mathbb{R}^n$ and assume that $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a function that is*

*(a) continuously differentiable in $x^*$* or

*(b) proper, convex and lower semicontinuous as well as Lipschitz continuous in a neighbourhood of $x^*$.*

*Moreover, let $\{x^k\} \subset \mathbb{R}^n$ converge to $x^*$, $\{d^k\} \subset \mathbb{R}^n$ converge to $d^*$, and $\{t_k\} \subset \mathbb{R}_+$ converge to 0. Then*

$$\psi'(x^*; d^*) = \lim_{k\to\infty} \frac{\psi(x^k + t_k d^k) - \psi(x^k)}{t_k}.$$

Note that the Lipschitz continuity of $\psi$ in assumption (b) holds by default, if $x^*$ is a point in the interior of $\operatorname{dom}\psi$.

The remaining part of this section deals with quadratic functions $f$, i.e. we consider the function

$$\psi(x) = f(x) + \varphi(x) = \frac{1}{2}\|Ax - b\|^2 + \varphi(x)$$

for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. This is an important and until now maybe the most investigated collection of functions considered for problem (1.1). Since $f$ is convex, every stationary point is a (global) minimizer of $\psi$. The next result considers the function values of these minimizers.

**Proposition 2.49.** *Let $f(x) := \frac{1}{2}\|Ax - b\|^2$ for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. If $x^*, y^* \in \mathbb{R}^n$ are minimizers of $\psi = f + \varphi$, there holds $Ax^* = Ay^*$. In particular, this yields $\varphi(x^*) = \varphi(y^*)$.*

*Proof.* Since $x^*$ is a minimizer of $\psi$, we use Lemma 2.46 and Proposition 2.47 to get

$$0 \in \partial\psi(x^*) = A^T(Ax^* - b) + \partial\varphi(x^*).$$

Thus, $-A^T(Ax^* - b)$ is a subgradient of $\varphi$ in $x^*$, which yields

$$\varphi(y^*) \geq \varphi(x^*) - \big(A^T(Ax^* - b)\big)^T(y^* - x^*).$$

Switching the roles of $x^*$ and $y^*$ there also holds

$$\varphi(x^*) \geq \varphi(y^*) - \big(A^T(Ay^* - b)\big)^T(x^* - y^*).$$

By adding both inequalities and simplifying, we obtain $\|A(x^* - y^*)\|^2 \leq 0$, and thus $Ax^* = Ay^*$, which yields $f(x^*) = f(y^*)$. Since both, $x^*$ and $y^*$ are minimizers of $\psi$, there holds $\psi(x^*) = \psi(y^*)$, hence also $\varphi(x^*) = \varphi(y^*)$, and this completes the proof. □

We close the section with a note on proximal methods for problem (1.1) if $f$ is quadratic.

**Remark 2.50.** Assume that we are interested in solving (1.1) with some proximal method, where the smooth function $f$ is quadratic, i.e. $f(x) := \frac{1}{2}\|Ax - b\|^2$ for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Since $\nabla^2 f(x) = A^T A$ is available in this situation, it is natural to consider the subproblem (1.2) using the second order information $H_k = A^T A$. Thus, the subproblem (1.2) is equivalent to

$$\arg\min_d \left\{ \frac{1}{2}\big(A(x^k + d) - b\big)^T \big(A(x^k + d) - b\big) + \varphi(x^k + d) \right\}$$

and thereby coincides with the initial problem. This implies, applying proximal methods with exact second order information (provided by the Hessian) to problems with quadratic

$f$ does not simplify the problem. Moreover, these methods terminate after solving only one subproblem with the precise minimizer of (1.1), assuming that it is possible to solve the subproblems exactly. ♢

## 2.3 Introduction to Kurdyka-Łojasiewicz-Functions

In the beginning of developing proximal methods for problems with composite functions, common assumptions for obtaining convergence and convergence rates were convexity or even strong convexity of $f$ and Lipschitz continuity of $\nabla f$. To overcome these requirements, recently error bound conditions or the Kurdyka-Łojasiewicz property have been widely used as assumptions for convergence proofs. We discuss the latter in this section and start with an outline of its historical progress.

Łojasiewicz [104] introduced a powerful condition to derive convergence results for gradient-type methods. That is, for a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ there exists $\rho \in [\frac{1}{2}, 1)$ such that the quantity

$$\frac{|f(x) - f(x^*)|^\rho}{\|\nabla f(x)\|} \tag{2.3}$$

remains bounded in a neighbourhood of any critical point $x^*$ of $f$. He proved that this property holds for any real-analytic function. On the other hand, it is known that the condition fails for some smooth (twice- or more times differentiable) functions.

**Example 2.51.** Let

$$f(x) = \begin{cases} \exp\left(-\frac{1}{x^2}\right), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0. \end{cases}$$

This function is well-known to be infinitely many times continuously differentiable and its only critical point is $x^* = 0$. The quantity (2.3) is equal to $\frac{|x|^3}{2} \exp\left(\frac{1-\rho}{x^2}\right)$, which is unbounded for $x \to 0$. Thus, the condition by Łojasiewicz does not hold. ♢

Kurdyka [90] generalized this idea to get the following property: There exists $\nu > 0$, a neighbourhood $U$ of a critical point $x^*$ and a continuous function $\phi : [0, \nu) \to \mathbb{R}_+$, which is continuously differentiable on $(0, \nu)$, there hold $\phi' > 0$ in $(0, \nu)$, $\phi(0) = 0$ and

$$\|\nabla(\phi \circ f)\| \geq 1 \tag{2.4}$$

for all $x \in U \cap [f(x^*) < f < f(x^*) + \nu]$, where $[a < f < b] := \{x \in \mathbb{R}^n : a < f(x) < b\}$. The Łojasiewicz inequality (2.3) is a special case of this definition with $\phi(s) = \left(s - f(x^*)\right)^{1-\rho}$. Inequality (2.4) holds for a wide range of functions, the ones definable on an o-minimal structure [90]. Note that inequality (2.4) is sometimes referred to as Kurdyka-Łojasiewicz inequality, but we will use this term for its nonsmooth variant.

For our purpose, we use a nonsmooth generalization of this condition, introduced in [22], but we use the formulation from [7]:

**Definition 2.52.** Let $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and lower semicontinuous. We say that $\psi$ has the *Kurdyka-Łojasiewicz (KL) property* in a point $\overline{x} \in \text{dom}(\partial\psi)$ if there exists $\nu > 0$, a neighbourhood $U$ of $\overline{x}$, and a continuous concave function $\phi : [0, \nu) \to \mathbb{R}_+$ such that

- $\phi(0) = 0$,
- $\phi$ is continuously differentiable for $s \in (0, \nu)$,
- $\phi'(s) > 0$ on $(0, \nu)$,

- for all $x \in U \cap [\psi(\overline{x}) < \psi < \psi(\overline{x}) + \nu]$, the *Kurdyka-Łojasiewicz (KL) inequality*

$$\phi'\big(\psi(x) - \psi(\overline{x})\big) \cdot \operatorname{dist}\big(0, \partial\psi(x)\big) \geq 1 \tag{2.5}$$

  holds, where $[a < \psi < b] := \{x \in \mathbb{R}^n : a < \psi(x) < b\}$.

If $\psi$ satisfies the KL-property at all points of its domain, $\psi$ is called *Kurdyka-Łojasiewicz (KL) function.*

Bolte, Danilidis and Shiota show in [22] that subanalytic functions and functions definable on o-minimal structures, so-called *tame* functions, are KL-functions. In particular, real polynomials, $p$-norms, exponential functions, and logarithms are examples of KL-functions. Furthermore, the class of tame functions is stable under finite sums and compositions, which yields a broad class of KL-functions. [7].
In the following, we provide a simple example and state results on KL-functions, which are necessary for our further analysis.

**Example 2.53.** Let $\psi(x) = x^2 + |x|$. Then $\psi$ has the KL-property in $\overline{x} = 0$ with $U = \mathbb{R}$, $\nu = +\infty$ and $\phi(x) = \sqrt{s}$. To prove the KL-inequality, note that Lemma 2.46 yields $\partial\psi(x) = \{2x + \operatorname{sign}(x)\}$ for $x \neq 0$. Thus, with $\psi(\overline{x}) = 0$, we get

$$\inf_{x \neq 0} \phi'\big(\psi(x)\big) \operatorname{dist}\big(0, \partial\psi(x)\big) = \inf_{x \neq 0} \frac{|2x + \operatorname{sign}(x)|}{2\sqrt{x^2 + |x|}} = 1,$$

and therefore the KL-inequality holds. ◊

In contrast to the earlier definition by Kurdyka, the function $\phi$ in Definition 2.52 is required to be concave. This is no major limitation, since for tame functions this property can be assumed without loss of generality [7]. On the other hand, unlike the previous formulations for smooth functions, the definition does not require $\overline{x}$ to be a stationary point of $\psi$. This restriction is superfluous, as the following result shows.

**Lemma 2.54.** *Let $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and lower semicontinuous. Then $\psi$ has the KL-property in any nonstationary point.*

For the proof, see Lemma 2 and Remark 4(b) in [7]. Moreover, a wide class of convex functions fulfils the KL-condition. Again, we refer to [7] and the references therein for the proofs.

**Proposition 2.55.** *Let $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, convex and lower semicontinuous function satisfying the following growth condition: If $x^*$ is a minimizer of $\psi$, then there exists a neighbourhood $U$ of $\overline{x}$, $\nu > 0$, $c > 0$ and $\rho \geq 1$ such that*

$$\psi(x) \geq \psi(x^*) + c \operatorname{dist}(x, \arg\min \psi)^\rho$$

*holds for all $x \in U \cap [\min \psi < \psi < \min f + \nu]$. Then $\psi$ has the KL-property in $x^*$ with $\phi(s) = \rho c^{-1/\rho} s^{1/\rho}$.*

For a strongly convex function $\psi$, this condition is trivially satisfied in its minimizer with $\rho = 2$, using that $0 \in \partial\psi(x^*)$ (Proposition 2.25) and Proposition 2.22, where $x^*$ is its (unique) minimizer.

**Corollary 2.56.** *Let $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lower semicontinuous and strongly convex with modulus $2\mu > 0$, i.e.*

$$\psi(y) \geq \psi(x) + d^T(y-x) + \mu\|y-x\|^2$$

*holds for all $x, y \in \operatorname{dom}\psi$ and $d \in \partial\psi(x)$. Then $\psi$ satisfies the KL-inequality with $\phi(s) = 2\mu^{-1/2}\sqrt{s}$.*

Since a large class of problems can be formulated using the KL-property, this has become a common assumption for proximal gradient-type algorithms in recent years, e.g. [7, 26, 64]. The advantage is that the property comprises much more types of functions than only strongly convex ones, but one can achieve similar convergence rates under this property. Nevertheless, the KL-property is no generalization of convexity. Bolte et al. [23] proved the existence of a convex, twice continuously differentiable function that does not fulfil the KL-property.

**Proposition 2.57.** *There exists a twice continuously differentiable and convex function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ which does not satisfy the KL-inequality and whose set of minimizers is compact with nonempty interior.*

They also showed that for all $k \in \mathbb{N}$ there exists a $k$-times continuously differentiable convex function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ which does not satisfy the KL-inequality. Hence, the KL-inequality is not related to the smoothness of $\varphi$.

**Remark 2.58.** Convergence proofs and proofs of convergence rates that assume the KL-property are often similar in structure. If $\{x^k\}$ is a sequence generated by some algorithm to minimize $\psi$, the analysis must include the following (or similar) statements as results or assumptions, cf. [64]:

- (Sufficient decrease) There exists $a > 0$ such that

$$\psi(x^{k+1}) - \psi(x^k) \leq -a\|x^{k+1} - x^k\| \quad \text{for all } k \geq 0.$$

- (Relative error) There exist $b > 0$ and a sequence $\{\nu_k\} \subset \mathbb{R}_+$ such that for all $k \geq 0$ there exists $s^{k+1} \in \partial\psi(x^{k+1})$ with

$$b\|s^{k+1}\| \leq \|x^{k+1} - x^k\| + \nu_{k+1} \quad \text{and} \quad \sum_{k=1}^{\infty} \nu_k < +\infty.$$

- (Continuity) There exists a subsequence $\mathcal{K} \subset \mathbb{N}_0$ such that $\{x^k\}_{\mathcal{K}}$ converges to some $x^*$ and $\{\psi(x^k)\}_{\mathcal{K}}$ converges to $\psi(x^*)$.

With these attributes and the KL-property one obtains that the sequence $\{x^k\}$ has finite length, i.e. $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| < +\infty$, converges to a stationary point of $\psi$ and if the KL-function is given by $\phi(s) = Cs^\theta$ for $C > 0$ and $\theta \in (\frac{1}{2}, 1]$, convergence rates can be obtained. Details of this procedure will be used in the local converge analysis of Algorithm 4.1. ♢

## 2.4 On Convergent Sequences

We close this chapter with some notes on the convergence of sequences. First, to clarify the used terminology, convergence rates are defined, before we state a result about isolated accumulation points.

The investigation of numerical algorithms is not complete without convergence rates, especially in the case of local convergence results. For this reason, the convergence rates required in the subsequent chapters are summarized below.

**Definition 2.59.** We say that a sequence $\{x^k\} \subset \mathbb{R}^n$ converges to $x^* \in \mathbb{R}^n$

(a) *linearly* or *Q-linearly*, if there exists $c \in (0,1)$ such that

$$\|x^{k+1} - x^*\| \leq c\|x^k - x^*\| \qquad \text{for all sufficiently large } k \geq 0,$$

(b) *superlinearly* or *Q-superlinearly*, if there exists a vanishing sequence $\{c_k\} \subset \mathbb{R}_+$ such that

$$\|x^{k+1} - x^*\| \leq c_k\|x^k - x^*\| \qquad \text{for all } k \geq 0,$$

(c) *quadratically* or *Q-quadratically*, if $\{x^k\}$ converges to $x^*$ and there exists $C > 0$ such that

$$\|x^{k+1} - x^*\| \leq C\|x^k - x^*\|^2 \qquad \text{for all } k \geq 0,$$

(d) *R-linearly*, *R-superlinearly* or *R-quadratically*, if there is a sequence $\{\varepsilon_k\} \subset \mathbb{R}_+$ converging to 0 Q-linearly, Q-superlinearly or Q-quadratically, respectively, and there holds

$$\|x^k - x^*\| \leq \varepsilon_k \qquad \text{for all } k \geq 0.$$

To end this section, we provide the following useful result, which considers isolated accumulation points of sequences. Recall that $x^*$ is an accumulation point of a sequence $\{x^k\}$, if there is a subsequence converging to $x^*$. The accumulation point $x^*$ is isolated, if the sequence has no further accumulation points in an appropriate neighbourhood of $x^*$. In this situation, the following holds, see [117, Lemma 4.10].

**Proposition 2.60.** *Let $x^* \in \mathbb{R}^n$ be an isolated accumulation point of the sequence $\{x^k\} \subset \mathbb{R}^n$. If for all $\mathcal{K} \subset \mathbb{N}_0$ such that $\{x^k\}_{\mathcal{K}}$ converges to $x^*$ we have $\{\|x^{k+1} - x^k\|\}_{\mathcal{K}} \to 0$, then the complete sequence converges to $x^*$.*

# Chapter 3

# The Proximity Operator and the Proximal Gradient Method

The proximity operator was introduced by Moreau [115,116] in the 1960s and was thereupon shown to be an important tool both theoretically and numerically. Subsequently, it became the base operation of numerous standard methods for composite optimization problems.
In the following, let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex, lower semicontinuous and proper function, $H \in \mathbb{S}^n_{++}$ and $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in a neighbourhood of $\operatorname{dom} \varphi$.
The *proximity operator* of $\varphi$ with respect to $H$ in a point $x \in \mathbb{R}^n$ is defined via

$$\operatorname{prox}^H_\varphi(x) := \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y - x\|^2_H + \varphi(y) \right\} = \operatorname*{arg\,min}_{y \in \operatorname{dom} \varphi} \left\{ \frac{1}{2}\|y - x\|^2_H + \varphi(y) \right\}. \tag{3.1}$$

Note that the objective function $y \mapsto \frac{1}{2}\|y - x\|^2_H + \varphi(y)$ is strongly convex. Hence, the proximity operator is single-valued in every point $x$. Thus, for simplicity, we refer to $\operatorname{prox}^H_\varphi(x)$ as an element in $\mathbb{R}^n$ instead of an one-element set. Moreover, we write

$$\operatorname{prox}_\varphi := \operatorname{prox}^I_\varphi,$$

if $H = I$.

For a closed, convex set $C \subset \mathbb{R}^n$, the proximity operator $\operatorname{prox}_{\iota_C}$ of the indicator function $\iota_C$ reduces to the classical orthogonal projection onto the set $C$. We demonstrate the imaging behaviour of $\operatorname{prox}_\varphi$ for an arbitrary function $\varphi$ in Figure 3.1, similar to [132]. In this example, the thin black lines are contour lines of $\varphi$ and the thick black line represents the boundary of $\operatorname{dom} \varphi$. The proximity operator maps the blue square points onto the green circles, whereas the red star represents the minimizer of $\varphi$. It can be observed that points in the domain are mapped towards the minimizer and points outside the domain are mapped onto the boundary.

Basic and useful properties of the proximity operator, which are required in the subsequent chapters, are summarized in Section 3.1. The following Section 3.2 deals with the proximal gradient method, which is a basic method for the solution of (1.1). After stating the algorithm, we provide the convergence theory under quite mild assumptions. Depending on the preliminaries, we also derive some results regarding the rate of convergence in Section 3.2.3.
One of the most difficult tasks in the context of proximal methods is the – theoretical and numerical – computation of the proximity operator and the solution of the underlying minimization problem. Although we will see in Section 3.1 that in many cases of interest

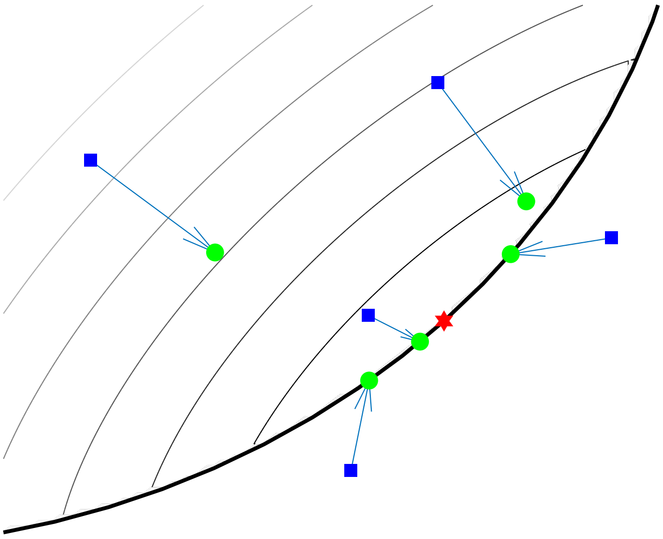


Figure 3.1: Example for the Evaluation of the Proximity Operator

there is an analytic representation of $\mathrm{prox}_{\varphi}$, the problem of computing $\mathrm{prox}_{\varphi}^{H}$ for some $H \neq I$ is no easy task and often impossible analytically. In Section 3.3 we deduce an algorithm for the computation of the proximity operator for special low rank matrices $H$.

## 3.1 Basic Properties of the Proximity Operator

This section is a collection of basic results regarding the proximity operator, supplemented by some examples. For a very detailed discussion about proximity operators we refer to the monograph by Bauschke and Combettes [12], whereas Beck [13] contains a large collection of specific formulas for computing the proximity operator of some functions $\varphi$. While not stated explicitly, the results presented in this section are taken from [13, Chapter 6].
We start with the separability of the proximity operator, which is thereafter used to compute the proximity operator of the $\ell_1$-norm.

**Proposition 3.1** (Separability of the Proximity Operator)**.** *Let $\varphi$ be separable, i.e. there exist convex, proper and lower semicontinous functions $\varphi_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ with positive integers $n_i$ $(i = 1, \dots, s)$ such that $n = n_1 + \dots + n_s$ and*

$$\varphi(x_1, \dots, x_s) = \sum_{i=1}^{s} \varphi_i(x_i).$$

*Then, for $x = (x_1, \dots, x_s)$, $\mathrm{prox}_{\varphi}(x) = \mathrm{prox}_{\varphi_1}(x_1) \times \dots \times \mathrm{prox}_{\varphi_s}(x_s)$.*

Maybe some of the most common convex, nonsmooth functions are the $\ell_1$-norm and the Euclidean norm. The following result provides formulas for computing the proximity operator of these functions.

**Proposition 3.2.** *Let $\lambda > 0$. Then*

$$\mathrm{prox}_{\lambda\|\cdot\|_1}(x) = \mathrm{sign}(x) \odot \max\{|x| - \lambda, 0\},$$

*where $\odot$ denotes the element-wise product and the application of the operations* sign *and* max *is also understood element-wise, and*

$$\operatorname{prox}_{\lambda\|\cdot\|_2}(x) = \left(1 - \frac{\lambda}{\max\{\|x\|_2, \lambda\}}\right) x.$$

We continue with a result, which essentially shows how to compute the proximity operator of the composition of a convex and some linear functions.

**Proposition 3.3.** *Let $A \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. Then*

$$\operatorname{prox}^H_{\varphi\circ A}(x) = A^T \operatorname{prox}^{AHA^T}_{\varphi}(Ax).$$

Note that it is possible to prove a more general result for general matrices $A \in \mathbb{R}^{m\times n}$ under appropriate assumptions, see Lemma 3.2.2 and Remark 3.2.3 in [111].
The remaining part of this section is devoted to providing some general results about proximity operators of convex, lower semicontinuous, proper functions. The first one is a very simple, but useful conclusion from Fermat's rule (Proposition 2.47) applied to the objective function in (3.1) and using some elementary transformations.

**Lemma 3.4.** *The following equivalences hold for $x, p \in \mathbb{R}^n$ and $H \in \mathbb{S}^n_{++}$:*

$$p = \operatorname{prox}^H_\varphi(x) \quad \Longleftrightarrow \quad H(x-p) \in \partial\varphi(p) \quad \Longleftrightarrow \quad p \in x - H^{-1}\partial\varphi(p).$$

As a first consequence of these characterizations, we prove the nonexpansiveness of the proximity operator, which entails continuity and Lipschitz continuity.

**Proposition 3.5.** *The proximity operator $\operatorname{prox}^H_\varphi$ is firmly nonexpansive with respect to the norm induced by $H$, i.e. for any $x, y \in \mathbb{R}^n$ we have*

$$\left\|\operatorname{prox}^H_\varphi(x) - \operatorname{prox}^H_\varphi(y)\right\|^2_H \le \left\langle \operatorname{prox}^H_\varphi(x) - \operatorname{prox}^H_\varphi(y), x - y\right\rangle_H .$$

*Proof.* This result was already shown by Moreau in [115], but the presented proof is taken from [50]. Let $p = \operatorname{prox}^H_\varphi(x)$ and $q = \operatorname{prox}^H_\varphi(y)$. Then Lemma 3.4 implies

$$\begin{aligned}\varphi(q) &\ge \varphi(p) + (x-p)^T H(q-p)\\ \varphi(p) &\ge \varphi(q) + (y-q)^T H(p-q).\end{aligned}$$

Adding both estimates yields the assertion. □

We conclude this section by noting that the more general mapping $(x, H) \mapsto \operatorname{prox}^H_\varphi(x)$ for $(x, H) \in \mathbb{R}^n \times \mathbb{S}^n_{++}$ is also continuous. Since this result is not needed in the following, we refer the interested reader to [111, Corollary 3.1.4] for the proof. Furthermore, we note that the result of Lemma 3.6 is closely related to that property.

## 3.2 The Proximal Gradient Method

The proximity operator was used by Rockafellar [140] to develop the proximal point method, an optimization method for convex functions. Based on this, Fukushima and Mine [66] introduced the basic form of the proximal gradient method for the solution of composite optimization problems in the form (1.1) in the 1980s. In the last decades, many variants of the method with different backtracking strategies to guarantee global convergence under

appropriate assumptions were introduced. One of the most popular ones is the Iterative Shrinkage Thresholding Method (ISTA), cf. Nesterov [120] and Beck and Teboulle [14], which opens up the possibility for several modifications, e.g. [8,164]. The method described in Tseng and Yun [156] uses a different approach with an Armijo-type line search. Similar methods are described by Milzarek [112] and Bonettini et al. [25,26]. Note that the latter mentioned class of methods does, in contrast to the group of methods based on ISTA, not require a (globally) Lipschitz continuous gradient of $f$.
In this section, we describe a basic proximal gradient algorithm similar to the one of Milzarek [112] with a classical line search, present the global convergence theory, and determine convergence rates under various assumptions.

### 3.2.1 Deduction of the Method

The following section mainly coincides with [82, Section 2.2]. To motivate the proximal gradient method, we derive it as a generalization of the classical (weighted) gradient method for the minimization of a smooth objective function $f : \mathbb{R}^n \to \mathbb{R}$. In this method, at step $k$ and at point $x^k$, the search direction $d^k$ is the solution of a minimization problem of the form

$$\min_d f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d$$

with some $H_k \in \mathbb{S}^n_{++}$, and the next iterate is $x^{k+1} = x^k + t_k d^k$ for some suitable step size $t_k > 0$. Usually, $H_k$ is chosen as a positive multiple of the identity matrix, since in this case the computation of the search direction is computationally inexpensive. For $H_k = I$, we get the method of steepest descent, as $d^k$ is given by $-\nabla f(x^k)$ in this case.
Next consider the nonsmooth optimization problem (1.1). To solve this problem, we again linearise the smooth part $f$ and add the nonsmooth function to obtain the subproblem

$$\min_d f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + \varphi(x^k + d) \tag{3.2}$$

as introduced in (1.2). Note that due to the positive definiteness of $H_k$ this problem has a unique solution, which is in the following denoted by $d^k$. Using the definition of the proximity operator in (3.1), a simple calculation shows that this solution is given by

$$d^k = \operatorname{prox}_\varphi^{H_k}\left(x^k - H_k^{-1}\nabla f(x^k)\right) - x^k. \tag{3.3}$$

The subsequent algorithm allows $H_k$ to be any positive definite matrix. In general, it is chosen independently of the iteration and as a positive multiple of the identity matrix, because in that case the computation of the proximity operator is less expensive computationally, in many relevant applications even an explicit expression is known, see also Section 3.1. However, we want to mention that there are several methods investigating the variation of $H_k$, referred to as variable metric proximal (gradient) methods.
The search direction $d^k$ depends on the choice of $H_k$. Nonetheless, search directions obtained from different matrices (but in the same iterate $x^k$) are related by the following result, which is essentially [156, Lemma 3]. A proof of the result is given in the appendix.

**Lemma 3.6.** *Let $x \in \mathbb{R}^n$ and $H, \widehat{H} \in \mathbb{S}^n_{++}$ be given and set*

$$d := \operatorname{prox}_\varphi^H\left(x - H^{-1}\nabla f(x)\right) - x, \qquad \widehat{d} := \operatorname{prox}_\varphi^{\widehat{H}}\left(x - \widehat{H}^{-1}\nabla f(x)\right) - x.$$

*Then*

$$\|\widehat{d}\| \leq \left(1 + \frac{\lambda_{\max}(\widehat{H})}{\lambda_{\min}(H)}\right) \cdot \frac{\lambda_{\max}(H)}{\lambda_{\min}(\widehat{H})} \cdot \|d\|.$$

Another important property of the search direction is that it vanishes if and only if $x^k$ is a stationary point of $\psi$ in the sense of Definition 2.43. Since we need a slightly more general result later on, it is formulated for arbitrary (not necessarily positive definite) matrices, which can be found in [81, Lemma 3.2].

**Proposition 3.7.** *Let $x^k \in \operatorname{dom}\varphi$, $H_k \in \mathbb{S}^n$ and*

$$d^k \in \operatorname*{arg\,min}_{d \in \mathbb{R}^n} \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + \varphi(x^k + d) \right\}. \tag{3.4}$$

*If $d^k = 0$, then $x^k$ is a stationary point of $\psi$. The converse is true if $H_k \in \mathbb{S}^n_{++}$.*

*Proof.* Assume that $d^k = 0$. From the definition of $d^k$ and Fermat's rule (Proposition 2.47), we get

$$0 \in \nabla f(x^k) + H_k d^k + \partial\varphi(x^k + d^k).$$

Setting $d^k = 0$ yields $0 \in \nabla f(x^k) + \partial\varphi(x^k)$, which is the desired result. Conversely, let $H_k$ be positive definite and $x^k$ a stationary point of $\psi$. Then $-\nabla f(x^k) \in \partial\varphi(x^k)$, which yields $\varphi(x^k + d) \geq \varphi(x^k) - \nabla f(x^k)^T d$ for any $d \in \mathbb{R}^n$. Thus,

$$\begin{aligned} \varphi(x^k) &\leq \nabla f(x^k)^T d + \varphi(x^k + d) \\ &< \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + \varphi(x^k + d) \end{aligned}$$

for any $d \neq 0$ and 0 is therefore the unique minimizer of the objective function in (3.4). □

It is easy to see that the converse statement of Proposition 3.7 may not hold if $H_k$ is not positive definite.
A simple consequence of this result, using the representation (3.3) of $d^k$, is the following characterization of stationary points of problem (1.1) in terms of a fixed point equation.

**Corollary 3.8.** *The point $x \in \mathbb{R}^n$ is a stationary point of $\psi$ if and only if*

$$\operatorname{prox}_\varphi^H \left(x - H^{-1}\nabla f(x)\right) = x$$

*for all $H \in \mathbb{S}^n_{++}$.*

To come back to the derivation of the proximal gradient method, after computing the search direction, the next iterate is defined by $x^{k+1} := x^k + t_k d^k$ for a suitable step size $t_k \in (0, 1]$. To compute this step size, we adapt a line search criterion from the classical Armijo line search, see e.g. [123] for more details. In particular, let $\Delta_k := \nabla f(x^k)^T d^k + \varphi(x^k + d^k) - \varphi(x^k)$. Then, we search for $t_k \in (0, 1]$ preferably large to satisfy

$$\psi(x^k + t_k d^k) \leq \psi(x^k) + t_k \sigma \Delta_k$$

for some $\sigma \in (0, 1)$. In contrast to the classical Armijo line search, the directional derivative $\psi'(x^k; d^k)$ is replaced by $\Delta_k$ here. This is because the calculation of the directional derivative for nondifferentiable functions is generally difficult, while the summands of $\Delta_k$ are inexpensive to calculate. Furthermore, a major drawback of the directional derivative of nondifferentiable functions is that it is not continuous, cf. Remark 2.29. The relation of $\Delta_k$ and the directional derivative $\psi'(x^k; d^k)$ is shown in the following result.

**Lemma 3.9.** *Let $x^k \in \operatorname{dom}\varphi$, $H_k \in \mathbb{S}^n_{++}$ and $d^k \in \mathbb{R}^n$ be defined by (3.3). Then the estimates*

$$\psi'(x^k; d^k) \leq \Delta_k \leq -(d^k)^T H_k d^k$$

*hold.*

*Proof.* Due to the monotonicity of the difference quotient of convex functions, cf. Proposition 2.28(a), we have

$$\varphi(x^k + d^k) - \varphi(x^k) \geq \frac{\varphi(x^k + td^k) - \varphi(x^k)}{t}$$

for any $t \in (0,1)$. Taking the limit $t \to 0$ yields $\varphi(x^k + d^k) - \varphi(x^k) \geq \varphi'(x^k; d^k)$ and adding $f'(x^k; d^k) = \nabla f(x^k)^T d^k$ proves the first inequality.
The proof of the second one is taken from [95, Proposition 2.4]. Since $d^k$ is the minimizer of (3.2), for any $t \in (0,1)$ there holds

$$\begin{aligned}
\nabla f(x^k)^T d^k + &\frac{1}{2}(d^k)^T H_k d^k + \varphi(x^k + d^k) \\
&\leq \nabla f(x^k)^T (td^k) + \frac{1}{2}t^2 (d^k)^T H_k d^k + \varphi(x^k + td^k) \\
&\leq t\nabla f(x^k)^T d^k + \frac{1}{2}t^2 (d^k) H_k d^k + t\varphi(x^k + d^k) + (1-t)\varphi(x^k),
\end{aligned}$$

where the last inequality uses the convexity of $\varphi$. By rearranging and simplifying terms we obtain

$$\begin{aligned}
&(1-t)\nabla f(x^k)^T d^k + \frac{1}{2}(1-t^2)(d^k)^T H_k d^k + (1-t)\big(\varphi(x^k + d^k) - \varphi(x^k)\big) \leq 0 \\
\iff\ &\nabla f(x^k)^T d^k + (1+t)\frac{1}{2}(d^k)^T H_k d^k + \varphi(x^k + d^k) - \varphi(x^k) \leq 0 \\
\iff\ &\nabla f(x^k)^T d^k + \varphi(x^k + d^k) - \varphi(x^k) \leq -(1+t)\frac{1}{2}(d^k)^T H_k d^k.
\end{aligned}$$

Taking the limit $t \to 1$ completes the proof. □

After this discussion of the properties of the proximal gradient method, the full method is presented in Algorithm 3.1, noting that the termination criterion in (S.2) is justified by Corollary 3.8.

**Algorithm 3.1** Proximal Gradient Method

```
(S.0) Choose $x^0 \in \operatorname{dom}\varphi$, $\beta, \sigma \in (0,1)$, and set $k := 0$.
(S.1) Choose $H_k \in \mathbb{S}^n_{++}$ and determine $d^k$ as the solution of (3.2).
(S.2) If $d^k = 0$: STOP.
(S.3) Compute $t_k = \max\{\beta^l : l = 0, 1, 2, \dots\}$ such that
          $\psi(x^k + t_k d^k) \leq \psi(x^k) + t_k \sigma \Delta_k,$          (3.5)
      where $\Delta_k := \nabla f(x^k)^T d^k + \varphi(x^k + d^k) - \varphi(x^k)$.
(S.4) Set $x^{k+1} := x^k + t_k d^k$, $k \leftarrow k + 1$, and go to (S.1).
```

We continue with the proof of well-definedness of the method in Algorithm 3.1. Note that in view of Corollary 3.8 the algorithm is well-defined, if the line search criterion in (3.5) is satisfied after finitely many steps.

**Proposition 3.10.** *Algorithm 3.1 is well-defined and a method of descent, meaning that* $\psi(x^{k+1}) < \psi(x^k)$ *holds for all* $k \geq 0$.

*Proof.* If in a fixed iteration $k$ the iterate $x^k$ is a stationary point, we have $d^k = 0$ by Proposition 3.7 and the algorithm terminates. Otherwise, there holds $d^k \neq 0$ and Lemma 3.9 yields $\Delta_k < 0$. Hence, we obtain by the first inequality in Lemma 3.9

$$\frac{\psi(x^k + td^k) - \psi(x^k)}{t} \leq \sigma\Delta_k$$

for all sufficiently small $t > 0$. By rearranging this inequality we see that the step size rule, and, consequently, the entire algorithm is well-defined. Furthermore, using $\Delta_k < 0$ in (S.3) yields

$$\psi(x^{k+1}) = \psi(x^k + t_k d^k) \leq \psi(x^k) + t_k\sigma\Delta_k < \psi(x^k).$$

□

### 3.2.2 Convergence Analysis

In this section we provide the theory for global convergence of the proximal gradient method. Although the final result is a special case of the convergence theorem in [156] and, moreover, carried out in [112], we provide the details here for two reasons. First, the proof shows once more that this algorithm is a generalization of the smooth gradient method, see e.g. [20, Proposition 1.2.1], and second, we need the statement of an intermediate auxiliary result of the proof in the further analysis.
For the convergence theory, we assume implicitly that the algorithm generates an infinite sequence of iterates $\{x^k\}$ and does not terminate after finitely many steps.

**Lemma 3.11.** *Let* $\{x^k\}$ *be a sequence such that* $x^{k+1} = x^k + t_k d^k$ *holds for all* $k \geq 0$ *with some search directions* $d^k \in \mathbb{R}^n$ *and* $t_k \in (0, 1]$. *Furthermore, assume that* $\psi(x^{k+1}) \leq \psi(x^k)$ *holds for all* $k \geq 0$. *Let* $\{x^k\}_\mathcal{K}$ *be a convergent subsequence of the given sequence such that the search direction* $d^k$ *is obtained from* (3.4) *for all* $k \in \mathcal{K}$, *where* $mI \preceq H_k \preceq MI$ $(0 < m \leq M)$ *holds, and the step size* $t_k \in (0, 1]$ *is determined by the Armijo-type rule* (3.5) *for all* $k \in \mathcal{K}$. *Then the limit point of* $\{x^k\}_\mathcal{K}$ *is a stationary point of* $\psi$ *and* $\{d^k\}_\mathcal{K} \to 0$.

*Proof.* Assume that $\{x^k\}_\mathcal{K}$ converges to $x^* \in \mathbb{R}^n$. By the lower semicontinuity of $\psi$ we have $\psi(x^*) \leq \liminf_{k\in\mathcal{K}, k\to\infty} \psi(x^k)$. Since, in addition, $\{\psi(x^k)\}$ is monotonically decreasing, the complete sequence $\{\psi(x^k)\}$ converges in $\mathbb{R}$. In particular $\left\{\psi(x^k) - \psi(x^{k+1})\right\}$ is a vanishing sequence. By Lemma 3.9 and the assumption $H_k \succeq mI$ we know that

$$\Delta_k \leq -(d^k)^T H_k d^k \leq -m\|d^k\|^2 \tag{3.6}$$

holds for $k \in \mathcal{K}$. Thus, we get from (3.5)

$$0 \leq \psi(x^k) - \psi(x^{k+1}) \leq \sigma t_k \Delta_k \leq -\sigma t_k m\|d^k\|^2 \leq 0,$$

which yields $\{t_k\|d^k\|^2\}_\mathcal{K} \to 0$.
Assume that the sequence $\{d^k\}_\mathcal{K}$ does not converge to 0. Then, by passing to a subsequence if necessary, there exists $\delta > 0$ such that $\|d^k\| \geq \delta$ holds for all $k \in \mathcal{K}$. By the above, this implies $\{t_k\}_\mathcal{K} \to 0$. Hence, for all sufficiently large $k \in \mathcal{K}$ we have $t_k \leq \beta$ and, by (3.5),

$$\psi\big(x^k + (t_k/\beta)d^k\big) - \psi(x^k) > \sigma(t_k/\beta)\Delta_k.$$

Rearranging terms yields

$$\begin{aligned}\sigma\Delta_k &< \frac{\psi\big(x^k + (t_k/\beta)d^k\big) - \psi(x^k)}{t_k/\beta}\\ &= \frac{f\big(x^k + (t_k/\beta)d^k\big) - f(x^k) + \varphi\big(x^k + (t_k/\beta)d^k\big) - \varphi(x^k)}{t_k/\beta}\\ &\le \frac{f\big(x^k + (t_k/\beta)d^k\big) - f(x^k) + (1 - t_k/\beta)\varphi(x^k) + (t_k/\beta)\varphi(x^k + d^k) - \varphi(x^k)}{t_k/\beta}\\ &= \frac{f\big(x^k + (t_k/\beta)d^k\big) - f(x^k)}{t_k/\beta} + \varphi(x^k + d^k) - \varphi(x^k)\\ &= \frac{f\big(x^k + (t_k/\beta)d^k\big) - f(x^k)}{t_k/\beta} - \nabla f(x^k)^T d^k + \Delta_k,\end{aligned}$$

where we used the convexity of $\varphi$ in the third line and the definition of $\Delta_k$ in the last one. Again, we rearrange terms and get

$$-(1-\sigma)\Delta_k < \frac{f\big(x^k + (t_k/\beta)d^k\big) - f(x^k)}{t_k/\beta} - \nabla f(x^k)^T d^k. \tag{3.7}$$

Inserting (3.6) into (3.7) and dividing the equation by $\|d^k\|$ yields

$$(1-\sigma)m\delta \le (1-\sigma)m\|d^k\| < \frac{f\big(x^k + (t_k/\beta \cdot \|d^k\|)d^k/\|d^k\|\big) - f(x^k)}{t_k/\beta \cdot \|d^k\|} - \frac{\nabla f(x^k)^T d^k}{\|d^k\|}.$$

As the sequence $\{t_k\|d^k\|^2\}_{\mathcal{K}}$ converges to 0 and $\|d^k\| \ge \delta$, we know that $t_k/\beta\|d^k\| \le t_k/(\beta\delta)\|d^k\|^2$, hence

$$\hat{t}_k := t_k/\beta \cdot \|d^k\| \to_{\mathcal{K}} 0.$$

By passing to a subsequence again, if necessary, we assume that $\{d^k/\|d^k\|\}$ converges to some $d^*$ in $\mathcal{K}$. Thus, taking the limit as $k \to \infty$, $k \in \mathcal{K}$ and using Proposition 2.48, we obtain

$$0 < (1-\sigma)m\delta \le \nabla f(x^*)^T d^* - \nabla f(x^*)^T d^* = 0,$$

which is a contradiction. Hence, we have $\{d^k\}_{\mathcal{K}} \to 0$.
For all $k \in \mathcal{K}$ and any $x \in \mathbb{R}^n$ the definition of $d^k$ yields

$$\nabla f(x^k)^T d^k + \frac{1}{2}(d^k)^T H_k d^k + \varphi(x^k + d^k) \le \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T H_k (x - x^k) + \varphi(x).$$

Since $\{H_k\}$ is bounded, by possibly passing to a subsequence, we assume that $\{H_k\}_{\mathcal{K}}$ converges to some $H \in \mathbb{S}^n_{++}$. Taking the limit (inferior) as $k \to \infty, k \in \mathcal{K}$, using the lower semicontinuity of $\varphi$ and the continuity of $f$ and $\nabla f$, we finally obtain

$$\varphi(x^*) \le \liminf_{\mathcal{K}\ni k \to \infty} \varphi(x^k + d^k) \le \nabla f(x^*)^T (x - x^*) + \frac{1}{2}(x - x^*)^T H (x - x^*) + \varphi(x)$$

for all $x \in \mathbb{R}^n$. Thus, 0 is a minimizer of the objective function in (3.2) and $x^*$ is a stationary point of $\psi$ by Proposition 3.7. □

The global convergence theorem for the proximal gradient method is a simple consequence of this result.

**Theorem 3.12.** *Let $\{H_k\}_k \subset \mathbb{S}^n_{++}$ be a sequence such that there exist $0 < m \leq M$ with $mI \preceq H_k \preceq MI$ for all $k \geq 0$. Then every accumulation point of a sequence generated by Algorithm 3.1 is a stationary point of $\psi$.*

### 3.2.3 Convergence Rates

In this section we analyse the asymptotic convergence rate of Algorithm 3.1 under appropriate conditions similar to the ones for smooth optimization [163]. It is convenient to assume Lipschitz continuity of $\nabla f$ to obtain results on convergence rates. Hence, we first prove further results under this condition. After that, we assume in addition that $f$ is a convex function. Finally, we state the linear convergence of the sequence of function values under strong convexity of $f$.

#### 3.2.3.1 Lipschitz Continuity of the Gradient of *f*

In the following analysis, let $\nabla f$ be Lipschitz continuous with Lipschitz constant $L > 0$, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad \text{for all } x, y \in \operatorname{dom}\varphi.$$

Then, the next result implies that the sequence of step sizes $\{t_k\}$ is bounded from below.

**Lemma 3.13.** *Suppose that $x^k \in \operatorname{dom}\varphi$ and $H_k \succeq mI$ for some $m > 0$. Then the estimate of the Armijo-type condition* (3.5) *is satisfied for all $t > 0$ with*

$$t \leq \min\left\{1, \frac{2m}{L}(1-\sigma)\right\}.$$

*Proof.* The proof is taken from [95]. Let $t \in (0,1]$ be arbitrary and $d^k$ the search direction in $x^k$ computed from (S.2) in Algorithm 3.1. Then

$$\begin{aligned}
\frac{\psi(x^k + td^k) - \psi(x^k)}{t} - \sigma\Delta_k &= \frac{f(x^k + td^k) - f(x^k)}{t} + \frac{\varphi(x^k + td^k) - \varphi(x^k)}{t} \\
&\quad - \left(\nabla f(x^k)^T d^k + \varphi(x^k + d^k) - \varphi(x^k)\right) + (1-\sigma)\Delta_k \\
&\leq \frac{f(x^k + td^k) - f(x^k)}{t} - \nabla f(x^k)^T d^k + (1-\sigma)\Delta_k \\
&\leq \int_0^1 \left(\nabla f(x^k + std^k) - \nabla f(x^k)\right)^T d^k \mathrm{d}s - (1-\sigma)(d^k)^T H_k d^k \\
&\leq \left(\frac{L}{2}t - m(1-\sigma)\right)\|d^k\|^2,
\end{aligned}$$

where the first inequality uses the monotonicity of the difference quotient of a convex function (Proposition 2.28(a)), Lemma 3.9 justifies the second one, and the last estimate follows from the Lipschitz continuity of $\nabla f$. Note that since $\operatorname{dom}\varphi$ is convex, $\nabla f$ is well defined in all points in the integral.
This estimate shows that the Armijo-type condition in (3.5) is met, whenever

$$\frac{L}{2}t - m(1-\sigma) \leq 0 \qquad \Longleftrightarrow \qquad t \leq \frac{2m}{L}(1-\sigma),$$

which completes the proof. □

Under the above assumptions, there is a lower bound for the step size $t_k$ in the proximal gradient method, namely

$$t_{\min} := \beta \cdot \min\left\{1, \frac{2m}{L}(1-\sigma)\right\}. \tag{3.8}$$

Using this lower bound we can prove a stronger convergence result in the case of Lipschitz continuous $\nabla f$.

**Theorem 3.14.** *Let $\{H_k\}_k \subset \mathbb{S}^n_{++}$ be a sequence such that there exist $0 < m$ with $mI \preceq H_k$ for all $k \geq 0$, $\nabla f$ be Lipschitz continuous and $\psi$ bounded from below. Then, for every sequence $\{x^k\}$ generated by Algorithm 3.1 the corresponding sequences $\{\Delta_k\}$ and $\{d^k\}$ converge to 0. If, in addition, there exists an isolated accumulation point $x^*$ of $\{x^k\}$, the complete sequence converges to $x^*$.*

*Proof.* By (3.8) there holds $t_k \geq t_{\min} > 0$ for all $k \geq 0$. Furthermore, since the sequence $\{\psi(x^k)\}$ is bounded from below and monotonically decreasing (Proposition 3.10), it converges to some $\psi^* \in \mathbb{R}$. Thus, the Armijo-type condition (3.5) yields

$$\psi(x^0) - \psi^* = \sum_{k=0}^{\infty} \psi(x^k) - \psi(x^{k+1}) \geq \sum_{k=0}^{\infty} -\sigma t_k \Delta_k \geq -\sigma t_{\min} \sum_{k=0}^{\infty} \Delta_k \geq 0,$$

where we used that $\Delta_k$ is nonpositive for all $k \geq 0$. This property immediately results in $\{\Delta_k\} \to 0$. Hence, Lemma 3.9 yields $\{d^k\} \to 0$.
Now, let $\{x^k\}_{\mathcal{K}}$ converge to $x^*$. The above yields

$$\|x^{k+1} - x^k\| = t_k\|d^k\| \leq \|d^k\| \to_{\mathcal{K}} 0.$$

Thus, Proposition 2.60 completes the proof. □

To conclude this section, we give an estimate on the size of $\|d^k\|$, as a generalization to the smooth result [163, Theorem 4.2.1] and note that this result does not require the sequence $\{H_k\}$ to be bounded (from above).

**Theorem 3.15.** *Let $\{H_k\}_k \subset \mathbb{S}^n_{++}$ be a sequence such that there exists $0 < m$ with $mI \preceq H_k$ for all $k \geq 0$, $\nabla f$ be Lipschitz continuous and $\psi$ bounded from below by $\psi^* \in \mathbb{R}$. Then, for the proximal gradient method in Algorithm 3.1 and any integer $\ell > 0$ we have*

$$\min_{0 \leq k \leq \ell-1} \|d^k\| \leq \sqrt{\frac{1}{\ell} \cdot \frac{1}{\sigma m t_{\min}} \big(\psi(x^0) - \psi(x^\ell)\big)} \leq \sqrt{\frac{1}{\ell} \cdot \frac{1}{\sigma m t_{\min}} \big(\psi(x^0) - \psi^*\big)}.$$

*Proof.* From (3.5), Lemma 3.9 and (3.8) we know

$$\psi(x^{k+1}) - \psi(x^k) \leq \sigma t_k \Delta_k \leq -\sigma t_k (d^k)^T H_k d^k \leq -\sigma t_{\min} m \|d^k\|^2.$$

Rearranging terms and summation yields

$$\sum_{k=0}^{\ell-1} \|d^k\|^2 \leq \sum_{k=0}^{\ell-1} \frac{1}{\sigma m t_{\min}} \big(\psi(x^k) - \psi(x^{k+1})\big) = \frac{1}{\sigma m t_{\min}} \big(\psi(x^0) - \psi(x^\ell)\big).$$

Since $\psi$ is bounded from below, we obtain

$$\begin{aligned}\min_{0\le k\le \ell-1}\|d^k\| = \sqrt{\min_{0\le k\le \ell-1}\|d^k\|^2} &\le \sqrt{\frac{1}{\ell}\sum_{k=0}^{\ell-1}\|d^k\|^2}\\ &\le \sqrt{\frac{1}{\ell}\cdot\frac{1}{\sigma m t_{\min}}\big(\psi(x^0)-\psi(x^\ell)\big)} \le \sqrt{\frac{1}{\ell}\cdot\frac{1}{\sigma m t_{\min}}\big(\psi(x^0)-\psi^*\big)}.\end{aligned}$$

□

#### 3.2.3.2 Convexity of *f*

If the function $f$ is, in addition, convex, we can prove a stronger estimate, which generalizes the corresponding result [163, Theorem 4.3.1] for smooth functions. In detail, let $x^*$ be a minimizer of $\psi$, then $\psi(x^k)-\psi(x^*)=\mathcal{O}(1/k)$, if $H_k=H$ is constant. We start with a technical lemma.

**Lemma 3.16.** *Suppose that $f$ is convex, and let the sequences $\{x^k\},\{d^k\},\{t_k\},\{H_k\}$ be defined as in Algorithm 3.1. Then for any $x\in\mathbb{R}^n$ and $k\ge 0$ there holds*

$$\psi(x)-\psi(x^k)\ge\frac{1}{\sigma t_k}\big(\psi(x^{k+1})-\psi(x^k)\big)+\frac{1}{2t_k}\big(\|x^{k+1}-x\|_{H_k}^2-\|x^k-x\|_{H_k}^2\big).$$

*Proof.* Fix $k\ge 0$. Since $f$ is convex, we have (Proposition 2.5)

$$f(x)-f(x^k)\ge\nabla f(x^k)^T(x-x^k). \tag{3.9}$$

Furthermore, set $y^k:=x^k+d^k$. Then, (3.5) and the definition of $\Delta_k$ yield

$$\begin{aligned}\psi(x^{k+1})-\psi(x^k)&\le\sigma t_k\big(\nabla f(x^k)^Td^k+\varphi(y^k)-\varphi(x^k)\big)\\ \iff\quad \varphi(y^k)-\varphi(x^k)&\ge-\nabla f(x^k)^Td^k+\frac{1}{\sigma t_k}\big(\psi(x^{k+1})-\psi(x^k)\big).\end{aligned} \tag{3.10}$$

Since $y^k=\operatorname{prox}_\varphi^{H_k}\big(x^k-H_k^{-1}\nabla f(x^k)\big)$, Lemma 3.4 yields

$$0\in H_k\big(y^k-x^k+H_k^{-1}\nabla f(x^k)\big)+\partial\varphi(y^k).$$

Thus, $-H_kd^k-\nabla f(x^k)\in\partial\varphi(y^k)$ and the definition of the subgradient (Definition 2.18) yield

$$\varphi(x)-\varphi(y^k)\ge-\big(H_kd^k+\nabla f(x^k)\big)^T(x-y^k) \tag{3.11}$$

for any $x\in\mathbb{R}^n$. Together, (3.9), (3.10), and (3.11) result in

$$\begin{aligned}\psi(x)-\psi(x^k)&=f(x)-f(x^k)+\varphi(x)-\varphi(y^k)+\varphi(y^k)-\varphi(x^k)\\ &\ge\frac{1}{\sigma t_k}\big(\psi(x^{k+1})-\psi(x^k)\big)-(d^k)^TH_k(x-y^k)\\ &=\frac{1}{\sigma t_k}\big(\psi(x^{k+1})-\psi(x^k)\big)+\|d^k\|_{H_k}^2-\frac{1}{t_k}(x^{k+1}-x^k)^TH_k(x-x^k).\end{aligned}$$

We use the identity

$$2(x^{k+1}-x^k)^TH_k(x-x^k)=\|x^{k+1}-x^k\|_{H_k}^2+\|x^k-x\|_{H_k}^2-\|x^{k+1}-x\|_{H_k}^2$$

and $x^{k+1} = x^k + t_k d^k$ to obtain

$$\begin{aligned}\psi(x) - \psi(x^k) &\geq \frac{1}{\sigma t_k}\big(\psi(x^{k+1}) - \psi(x^k)\big) + \|d^k\|_{H_k}^2 \\ &\quad - \frac{1}{2t_k}\big(\|x^{k+1} - x^k\|_{H_k}^2 + \|x^k - x\|_{H_k}^2 - \|x^{k+1} - x\|_{H_k}^2\big) \\ &= \frac{1}{\sigma t_k}\big(\psi(x^{k+1}) - \psi(x^k)\big) + \|d^k\|_{H_k}^2 \left(1 - \frac{t_k}{2}\right) \\ &\quad + \frac{1}{2t_k}\big(\|x^{k+1} - x\|_{H_k}^2 - \|x^k - x\|_{H_k}^2\big) \\ &\geq \frac{1}{\sigma t_k}\big(\psi(x^{k+1}) - \psi(x^k)\big) + \frac{1}{2t_k}\big(\|x^{k+1} - x\|_{H_k}^2 - \|x^k - x\|_{H_k}^2\big),\end{aligned}$$

where the final estimate uses $t_k \leq 1$. □

**Theorem 3.17.** *Suppose that $f$ is convex, $\nabla f$ is Lipschitz continuous with Lipschitz constant $L > 0$, $\psi$ is bounded from below, $x^* \in \arg\min \psi$ and $H_k = H \in \mathbb{S}^n_{++}$ for all $k \geq 0$. Then Algorithm 3.1 generates a sequence $\{x^k\}$ that satisfies*

$$\psi(x^k) - \psi(x^*) \leq \frac{1}{k} \cdot \frac{\frac{1}{\sigma}\big(\psi(x^0) - \psi(x^*)\big) + \frac{1}{2}\|x_0 - x^*\|_H^2}{t_{\min}}$$

*for all $k \geq 0$, where $t_{\min}$ is defined in* (3.8).

*Proof.* Lemma 3.16 and (3.8) yield

$$\begin{aligned}0 \geq \psi(x^*) - \psi(x^k) &\geq \frac{1}{\sigma t_k}\big(\psi(x^{k+1}) - \psi(x^k)\big) + \frac{1}{2t_k}\big(\|x^{k+1} - x^*\|_H^2 - \|x^k - x^*\|_H^2\big) \\ &\geq \frac{1}{\sigma t_{\min}}\big(\psi(x^{k+1}) - \psi(x^k)\big) + \frac{\gamma}{2t_{\min}}\big(\|x^{k+1} - x^*\|_H^2 - \|x^k - x^*\|_H^2\big),\end{aligned}$$

where the second estimate follows from the fact that the right hand side is nonpositive. Summing from $\ell = 0$ to $k$, we obtain

$$\begin{aligned}&(\ell + 1)\psi(x^*) - \sum_{k=0}^{\ell} \psi(x^k) \\ &\qquad \geq \frac{1}{\sigma t_{\min}}\big(\psi(x^{\ell+1}) - \psi(x^0)\big) + \frac{1}{2t_{\min}}\big(\|x^{\ell+1} - x^*\|_H^2 - \|x^0 - x^*\|_H^2\big) \\ &\qquad \geq \frac{1}{\sigma t_{\min}}\big(\psi(x^{\ell+1}) - \psi(x^0)\big) - \frac{1}{2t_{\min}}\|x^0 - x^*\|_H^2. \end{aligned} \tag{3.12}$$

Furthermore, from $\psi(x^{k+1}) - \psi(x^k) < 0$ we get

$$\begin{aligned}0 &\leq \sum_{k=0}^{\ell}(k+1)\big(\psi(x^k) - \psi(x^{k+1})\big) \\ &= \sum_{k=0}^{\ell} \psi(x^k) + \sum_{k=0}^{\ell} k \cdot \psi(x^k) - \sum_{k=1}^{\ell+1} k \cdot \psi(x^k) \\ &= \sum_{k=0}^{\ell} \psi(x^k) - (\ell + 1)\psi(x^{\ell+1}).\end{aligned}$$

Together with (3.12), we finally obtain

$$(\ell+1)\big(\psi(x^*)-\psi(x^{\ell+1})\big) \geq \frac{1}{\sigma t_{\min}}\big(\psi(x^*)-\psi(x^0)\big) - \frac{1}{2t_{\min}}\|x^0-x\|_H^2.$$

Rearranging terms yields the assertion. □

The assumption $H_k = H$ in Theorem 3.17 seems quite restrictive, however, it is common to obtain convergence rates for first-order methods. Note that the proof of Theorem 3.17 cannot simply be applied to a bounded sequence $\{H_k\}$ such that $mI \preceq H_k \preceq MI$ holds with some $0 < m \leq M$. The crucial point is (3.12), using the telescoping behaviour of $\|x^{k+1}-x^*\|_H - \|x^k - x^*\|_H$. We remark that a similar, but slightly more general result without this restriction was obtained in [25].
The resulting convergence rate $\mathcal{O}(1/k)$ is the same as for the classical Iterative Shrinkage Thresholding Algorithm, cf. [14]. However, numerical experiments of similar methods show, that the numerical performance of proximal gradient methods similar to the one stated in Algorithm 3.1 is even comparable to the accelerated version FISTA in [14] with a complexity of $\mathcal{O}(1/k^2)$, see e.g. [26] for a discussion.

#### 3.2.3.3 Strong Convexity of *f*

Under the assumption of strong convexity of $f$ it is even possible to prove linear convergence of the sequence $\{\psi(x^k)\}$. The result is a special case of a more general theorem stated by Tseng and Yun in [156] and the proof is given in the unpublished report [155]. For the sake of completeness, we provide details of the simplified proof in Appendix A.1.

**Theorem 3.18.** *Suppose that $f$ is strongly convex, $\nabla f$ is Lipschitz continuous, and $\psi$ is bounded from below with $x^* \in \arg\min \psi$. Furthermore let $0 < m \leq M$ such that $mI \preceq H_k \preceq MI$ holds for all $k \geq 0$. Then Algorithm 3.1 generates a sequence $\{x^k\}$ that satisfies*

$$\psi(x^{k+1}) - \psi(x^*) \leq c_1\big(\psi(x^k) - \psi(x^*)\big),$$

*for all $k \geq 0$, where $c_1 \in (0,1)$ is a constant depending on the Lipschitz constant of $\nabla f$, the strong convexity modulus of $f$, $m$, $M$ and the constants $\sigma$, $\beta$ of the Armijo-type line search.*

## 3.3 Numerical Computation of the Proximal Operator

One of the most important tasks in proximal algorithms is the efficient solution of the subproblem (3.2) and, equivalently, the computation of

$$\operatorname{prox}_\varphi^H(x - H^{-1}\nabla f(x)) \tag{3.13}$$

for $x \in \mathbb{R}^n$ and $H \in \mathbb{S}^n_{++}$. In some cases, like various norms, see Proposition 3.2, and other special functions, cf. [13], the computation of $\operatorname{prox}_\varphi$ is analytically possible. Due to this advantage, many first-order proximal methods use only multiples of the identity matrix, or in the case of separable functions $\varphi$ diagonal matrices for $H$.
On the other hand, often significantly better convergence properties are obtained if $H$ is chosen variable or even contains second-order information, i.e. corresponds to an approximation to the Hessian of $f$ in the current iterate. However, in that case it may not be possible to determine the proximity operator (3.13) analytically, or only with high effort. For that reason, it is eminently relevant to investigate efficient methods for numerically evaluating

the proximity operator. A general idea for this purpose is to use forward-backward splitting in a first-order proximal method [38, 95]. Other possibilities include fixed point [41] or interior point methods [65].
If $H$ has a specific structure, in particular for low-rank modifications of simple matrices, there are further efficient ways to compute proximity operators [16,84,145], at least inexactly. Here, we focus on a basic result of Becker, Fadili and Ochs in [15, 16], which was used there, however, only for rank-1 modifications. The combination of this approach with the compact representation of limited memory matrices from Byrd, Nocedal and Schnabel in [39] allows the development of a highly efficient method for computing proximity operators for matrices $H$ that are low-rank modifications of simple matrices such as the identity matrix. Our idea of this crucial combination makes [15, 16] applicable for numerical problems using a memory larger than one. The following elaboration is essentially based on [81], where the combination with compact representations for such matrices enables a large field of application. A related method for *identity minus rank one* matrices is given in [84].
First, we recall the details of limited memory quasi-Newton methods and especially the compact representation of the underlying matrices in Section 3.3.1. After that, the result for the computation of the proximity operator from [16] is introduced in Section 3.3.2, including notes on its consequences. Finally, we combine these techniques to get a highly efficient method for the numerical computation of proximity operators in Section 3.3.3.

### 3.3.1 Compact Representation of Limited Memory Quasi-Newton Matrices

The idea of using quasi-Newton matrices originates from smooth (quasi-)Newton methods using the recursion $x^{k+1} = x^k - H_k^{-1}\nabla f(x^k)$, but in fact, the approach is equally useful for proximal methods. For this we consider the problem of solving (3.13) in a point $x^k \in \mathbb{R}^n$. To obtain the best possible convergence properties, $H_k = \nabla^2 f(x^k)$ should hold. However, the major drawback of applying the Hessian is that it has no sparse structure in general and is therefore expensive to compute. The idea of quasi-Newton matrices is to use only an approximation to the Hessian and this turned out to be one of the best ideas to advance numerical optimization, cf. [123], as it can improve the performance significantly.
To the author's knowledge, there exist only few publications dealing with limited memory matrices and the advantages of their compact representation in combination with proximal-type methods, e.g. [84, 91]. The conjunction with the results in [16] outlines the benefits and makes this technique applicable to a wide class of applications, especially for large scale problems.

Defining

$$s^k := x^{k+1} - x^k \qquad \text{and} \qquad y^k := \nabla f(x^{k+1}) - \nabla f(x^k),$$

an approximation $H_{k+1} \approx \nabla^2 f(x^{k+1})$ has to satisfy the following basic properties: It should be symmetric and satisfy the *quasi-Newton equation* $H_{k+1}s^k = y^k$. Furthermore, it should be easy to calculate $H_{k+1}$ from $H_k$. Several methods were developed based on these assumptions, such as the SR1-, BFGS- and DFP-method and the Broyden class [123]. We focus on the well-known symmetric-rank-1-formula (SR1) and the rank-2-method from

Broyden, Fletcher, Goldfarb and Shanno (BFGS), which are given by the update formulas

$$H_{k+1}^{SR1} = H_k + \frac{(y^k - H_k s^k)(y^k - H_k s^k)}{(y^k - H_k s^k)^T s^k}, \tag{3.14}$$

$$H_{k+1}^{BFGS} = H_k + \frac{y^k (y^k)^T}{(s^k)^T y^k} - \frac{H_k s^k (s^k)^T H_k}{(s^k)^T H_k s^k}, \tag{3.15}$$

where $H_0$ is usually chosen to have simple structure. We note that the methods are well-defined as long as the denominators do not vanish. The advantage of the BFGS-update is that it maintains positive definiteness, as long as $(s^k)^T y^k > 0$ holds. In contrast, the computations in the SR1-update are cheaper, since the rank increases only by one in each step.

However, the benefits of the quasi-Newton updates shrink after several steps, since the rank of the approximation $H_k$ increases in each step. Hence, these quasi-Newton methods are not applicable to large-scale problems. We can avoid this problem by re-computing the matrix $H_{k+1}$ from $H_0$ and terms based on the vectors $s^j$ and $y^j$ for $j = k - m + 1, \ldots, k$ for some $m > 0$ instead of using all information for $j = 0, \ldots, k$. This means that we skip the first pairs of vectors $(s^j, y^j)$ and use only the $m$ most recent ones, which leads to limited memory quasi-Newton methods with memory $m$, see the fundamental work of Nocedal [122]. These limited memory versions may not start with the same initial matrix $H_0$, instead they often use an initialization $H_{k,0}$ depending on the current iterate $k$.

The crucial achievement for the numerical application of such methods lies in the compact representation of these matrices developed by Byrd, Nocedal and Schnabel [39]. They proved that many limited memory matrices have a representation of the form

$$H_k = H_{k,0} + A_k Q_k^{-1} A_k^T, \tag{3.16}$$

where $H_{k,0} \in \mathbb{S}_{++}^n$ is the initialization matrix, $Q_k \in \mathbb{R}^{s \times s}$ is symmetric and nonsingular and $A_k \in \mathbb{R}^{n \times s}$ with $s \ll n$. We provide the corresponding results for the SR1- and BFGS-update in the following. For that purpose, assume that a sequence $\{x^k\}$ is given such that $\{s^k\}$ and $\{y^k\}$ can be computed, and $m > 0$ is a fixed integer. For simplicity in the notation, assume that $k \geq m$. Then we define

$$S_k := [s^{k-m} \ldots s^{k-1}] \in \mathbb{R}^{n \times m} \qquad \text{and} \qquad Y_k := [y^{k-m} \ldots y^{k-1}] \in \mathbb{R}^{n \times m}.$$

Furthermore, let $D_k := D(S_k^T Y_k)$ and $L_k := L(S_k^T Y_k)$ denote the diagonal part and the strict lower triangle of the matrix $S_k^T Y_k$. Then we obtain the following results, where we skip the elementary proofs and refer the interested reader to the corresponding Theorems 2.3 and 5.1 in [39].

**Theorem 3.19.** *Let $H_{k,0} \in \mathbb{S}_{++}^n$ and let $H_k$ be obtained by updating $H_{k,0}$ $m$ times using the SR1-formula* (3.14) *and the pairs $\{s^j, y^j\}$. Assume that each update is well-defined, i.e. $(y^j - H_j s^j)^T s^j \neq 0$ for $j = k - m, \ldots, k - 1$. Then*

$$H_k = H_k^{SR1} = H_{k,0} + (Y_k - H_{k,0} S_k)(D_k + L_k + L_k^T - S_k^T H_{k,0} S_k)^{-1} (Y_k - H_{k,0} S_k)^T,$$

*and the matrix $D_k + L_k + L_k^T - S_k^T H_{k,0} S_k$ is nonsingular.*

Hence, in the notation of (3.16) we have

$$A_k = Y_k - H_{k,0} S_k \in \mathbb{R}^{n \times m} \qquad \text{and} \qquad Q_k = D_k + L_k + L_k^T - S_k^T H_{k,0} S_k \in \mathbb{R}^{m \times m}.$$

**Theorem 3.20.** *Let $H_{k,0} \in \mathbb{S}^n_{++}$ and assume that the $m$ pairs $\{s^j, y^j\}$ for $j = k-m, \dots, k-1$ satisfy $(s^j)^T y^j > 0$. Let $H_k$ be obtained by updating $H_{k,0}$ $m$ times using the BFGS-formula* (3.15) *and the pairs $\{s^j, y^j\}$. Then*

$$H_k = H_k^{BFGS} = H_{k,0} - \begin{bmatrix} H_{k,0}S_k & Y_k \end{bmatrix} \begin{bmatrix} S_k^T H_{k,0} S_k & L_k \\ L_k^T & -D_k \end{bmatrix}^{-1} \begin{bmatrix} S_k^T H_{k,0} \\ Y_k^T \end{bmatrix}.$$

This means, with the notation from (3.16) we get

$$A_k = \begin{bmatrix} H_{k,0}S_k & Y_k \end{bmatrix} \in \mathbb{R}^{n\times 2m} \qquad \text{and} \qquad Q_k = \begin{bmatrix} -S_k^T H_{k,0} S_k & -L_k \\ -L_k^T & D_k \end{bmatrix} \in \mathbb{R}^{2m\times 2m}.$$

Let us mention that [39] contains detailed descriptions of how to compute matrix-vector-products including the matrix $H_k$ efficiently from these representations.

### 3.3.2 Reduction of the Proximity Computation to a Small-Dimensional Semismooth System of Equations

In the following we show that the compact representation of limited memory matrices enables us to reduce the minimization problem (3.2) with dimension $n$ to a semismooth Newton system of dimension $m$ or $2m$, depending on the selected limited memory quasi-Newton method. Since $m$ is usually chosen very small, this reduces the complexity of the problem significantly.

The idea is to rewrite (3.16) in the form

$$H = H_0 + U_1 U_1^T - U_2 U_2^T \tag{3.17}$$

with suitable matrices $U_i \in \mathbb{R}^{n\times r_i}$ for small $r_i > 0$ $(i = 1, 2)$ and a simple matrix $H_0$ (typically a multiple of the identity matrix such that the corresponding proximal subproblem is easy to solve), so that $H$ is obtained from $H_0$ by a small rank modification. Note that, to simplify notation, we omit the dependence of these matrices from $k$ in this section. It is trivial to see that a single update in (3.14) or (3.15) is precisely of the form required in (3.17). However, since the additive terms in these quasi-Newton updates depend on $H_k$ itself, these formulas can not be used numerically (directly) to get this form for a memory larger than one without additional computation costs in every iteration. Using the compact representation (3.16) instead, this is tackled with little effort.

To this end, we compute a spectral decomposition $Q = V\Lambda V^T$ of $Q$, i.e. $V \in \mathbb{R}^{s\times s}$ is orthogonal and $\Lambda \in \mathbb{R}^{s\times s}$ is a diagonal matrix with diagonal entries $\lambda_i$. Note that the computation of this spectral decomposition is not time consuming since $s$ is small. We then define the sets $\mathcal{I}_+ := \{i : \lambda_i > 0\}$ and $\mathcal{I}_- := \{i : \lambda_i < 0\}$ to split $\Lambda$ into its positive and negative part

$$\Lambda = \begin{pmatrix} \Lambda_+ & 0 \\ 0 & -\Lambda_- \end{pmatrix}$$

with $\Lambda_+ = \Lambda_{[\mathcal{I}_+\mathcal{I}_+]}$ and $\Lambda_- = -\Lambda_{[\mathcal{I}_-\mathcal{I}_-]}$. The set $\mathcal{I}_0 := \{i : \lambda_i = 0\}$ can be omitted in the following analysis, even though it might be nonempty (note that this cannot happen for the BFGS-update, if $(s^k)^T y^k > 0$ holds always). The resulting matrices $\Lambda_+, \Lambda_-$ are positive definite and therefore have a matrix square root and an inverse. Defining

$$U_1 := (AV)_{[\cdot\mathcal{I}_+]}\Lambda_+^{-1/2} \qquad \text{and} \qquad U_2 := (AV)_{[\cdot\mathcal{I}_-]}\Lambda_-^{-1/2}$$

yields the representation in (3.17). With this transformation of the limited memory quasi-

Newton matrix we can apply a result by Becker, Fadili, Ochs [16, Corollary 3.6] to reduce the computation of the proximity operator $\operatorname{prox}_\varphi^H$, which is an $n$-dimensional minimization problem, to an $(r_1 + r_2)$-dimensional semismooth Newton system with $r_1 + r_2 \leq s$.

**Theorem 3.21.** *Let $H = H_0 + U_1U_1^T - U_2U_2^T \in \mathbb{S}_{++}^n$ with $H_0 \in \mathbb{S}_{++}^n$ and $U_i \in \mathbb{R}^{n\times r_i}$ with rank $r_i$ $(i = 1, 2)$. Set $H_1 = H_0 + U_1U_1^T$. Then, the following holds:*

$$\operatorname{prox}_\varphi^H(y) = \operatorname{prox}_\varphi^{H_0}(y + H_1^{-1}U_2\alpha_2^* - H_0^{-1}U_1\alpha_1^*), \tag{3.18}$$

*where $\alpha_i^* \in \mathbb{R}^{r_i}$, $i = 1, 2$, are the unique zeros of the coupled system $F(\alpha) = F(\alpha_1, \alpha_2) = 0$, where $F = \big(F_1, F_2\big)$ is defined via*

$$\begin{aligned} F_1(\alpha_1, \alpha_2) &= U_1^T(y + H_1^{-1}U_2\alpha_2 - \operatorname{prox}_\varphi^{H_0}(y + H_1^{-1}U_2\alpha_2 - H_0^{-1}U_1\alpha_1)) + \alpha_1, \\ F_2(\alpha_2, \alpha_2) &= U_2^T(y - \operatorname{prox}_\varphi^{H_0}(y + H_1^{-1}U_2\alpha_2 - H_0^{-1}U_1\alpha_1)) + \alpha_2. \end{aligned} \tag{3.19}$$

We skip the proof of the result, as this is an immediate consequence of applying [16, Theorem 3.4] twice, which is an easier version of the result for $H = H_0 + U_1U_1^T$ or $H = H_0 - U_2U_2^T$. The mapping $F$ is indeed Newton differentiable, as we see in the next section, and the corresponding system of equations $F(\alpha) = 0$ can therefore be solved with standard semismooth Newton solvers such as the one in Algorithm 2.1, which reduces the computation costs significantly.

### 3.3.3 The Full Algorithm

In this section we exploit the ideas of the previous sections and combine them with some further details to finally obtain the method for the computation of (3.13). Since the proximity operator is Lipschitz continuous (Proposition 3.5), a Newton derivative exists (Proposition 2.38) and semismooth Newton methods are suitable candidates for the numerical computation of the unique zero $\alpha^* = (\alpha_1^*, \alpha_2^*)$ of the nonlinear system of equations $F(\alpha) = 0$ in Theorem 3.21. An iteration of the semismooth Newton method is given by

$$\alpha^{j+1} = \alpha^j - G_j^{-1}F(\alpha^j),$$

where $G_j = G(\alpha_j)$ is a Newton derivative of $F$ in $\alpha^j$, cf. Algorithm 2.1. Provided that the Newton derivative of the proximity operator can be evaluated (analytically or using an efficient numerical method), a short calculation and the chain rule for Newton derivatives (Proposition 2.41) show the following result.

**Proposition 3.22.** *Let $\operatorname{prox}_\varphi^{H_0}$ be Newton differentiable with generalized derivative $P$. Then $F$ (as defined in Theorem 3.14) is Newton differentiable, and a generalized derivative is given by*

$$G(\alpha) = \begin{bmatrix} U_1 & U_2 \end{bmatrix}^T P(z) \begin{bmatrix} H_0^{-1}U_1 & -H_1^{-1}U_2 \end{bmatrix} + \begin{bmatrix} I & U_1^TH_1^{-1}U_2 \\ 0 & I \end{bmatrix},$$

*where $z = y + H_1^{-1}U_2\alpha_2 - H_0^{-1}U_1\alpha_1$.*

In many applications it is possible to compute the generalized derivative of the proximity operator analytically, as long as $H_0$ is a positive multiple of the identity matrix. Examples are the very common $\ell_1$- and $\ell_2$-norm.

**Example 3.23.** We use the formulas for the proximity operators from Proposition 3.2 to see that the diagonal matrix $P(x)$ with diagonal entries

$$P_{[ii]}(x) = \begin{cases} 1, & \text{if } |x_i| \geq \lambda\gamma, \\ 0, & \text{otherwise} \end{cases}$$

is an element of the generalized Jacobian in the sense of Clarke, and, therefore, a Newton derivative of $\operatorname{prox}^{1/\gamma I}_{\lambda\|\cdot\|_1}$. A short calculation shows further that

$$\hat{P}(x) = \begin{cases} \big(1 - \frac{\lambda\gamma}{\|x\|_2}\big)I + \frac{\lambda\gamma}{\|x\|_2^3} xx^T, & \text{if } \|x\|_2 \geq \lambda\gamma, \\ 0, & \text{otherwise.} \end{cases}$$

is a Newton derivative of $\operatorname{prox}^{1/\gamma I}_{\lambda\|\cdot\|_2}$. ◊

We summarize the previous discussion and present our method for the computation of a solution of (3.2), where $H$ is obtained using a limited memory quasi-Newton update, in Algorithm 3.2.

**Algorithm 3.2** Computation of the Proximity Operator with Respect to a Limited Memory Quasi-Newton Matrix

(S.0) Given an iterate $x \in \mathbb{R}^n$, a compact representation $H = H_0 + AQA^T$ of the corresponding Hessian approximation and a convex, proper, lower semicontinuous function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$.

(S.1) Compute the spectral decomposition $Q = V\Lambda V^T$ of $Q$, define $\mathcal{I}_+ := \{i : \lambda_i > 0\}$ and $\mathcal{I}_- := \{i : \lambda_i < 0\}$, set $\Lambda_+ = \Lambda_{[\mathcal{I}_+\mathcal{I}_+]}$ and $\Lambda_- = -\Lambda_{[\mathcal{I}_-\mathcal{I}_-]}$ and compute

$$U_1 := (AV)_{[\cdot\mathcal{I}_+]}\Lambda_+^{-1/2} \qquad \text{and} \qquad U_2 := (AV)_{[\cdot\mathcal{I}_-]}\Lambda_-^{-1/2}.$$

(S.2) Set $y = x - H^{-1}\nabla f(x)$, $H_1 = H_0 + U_1U_1^T$ and compute

$$H_1^{-1} = H_0^{-1} + H_0^{-1}U_1(I - U_1^TH_0^{-1}U_1)^{-1}U_1^TH_0^{-1}.$$

(S.3) Use a semismooth Newton method to determine the zero $\alpha^*$ of $F = (F_1, F_2)$ defined in (3.19) with a suitable termination criterion. In particular, apply the update rule

$$\alpha^{j+1} = \alpha^j - G(\alpha^j)^{-1}F(\alpha^j),$$

where $G$ is the Newton derivative of $F$ according to Proposition 3.22.

(S.4) Compute $d = \operatorname{prox}^H_\varphi(y) - x$ using (3.18).

Note that the formula for $H_1^{-1}$ in (S.2) is obtained using the Sherman-Morrison-Woodbury-identity. This formula could also be used to compute $H^{-1}$, but the direct application of this formula to the compact representation (3.16) saves us from several matrix-products including the (possibly full) matrix $H_1^{-1}$. Furthermore, we mention that an efficient computation of the product $H^{-1}\nabla f(x)$ is possible using the techniques in [39].

# Chapter 4

# A Globalized Inexact Proximal Newton-type Method

In this chapter we continue to consider the composite optimization problem (1.1), where $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex, proper and lower semicontinuous and $f : \mathbb{R}^n \to \mathbb{R}$ is (twice) continuously differentiable on an open set containing $\operatorname{dom} \varphi$. In contrast to the proximal gradient method in the previous chapter, proximal Newton and quasi-Newton methods make use of second order information of $f$ in the matrix $H_k$ when solving the subproblems

$$\arg\min_{d} \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + \varphi(x^k + d) \right\} \tag{4.1}$$

in a given iterate $x^k \in \mathbb{R}^n$. A proximal Newton method uses $H_k = \nabla^2 f(x^k)$, whereas only an approximation $H_k \approx \nabla^2 f(x^k)$ is chosen in proximal quasi-Newton methods. The advantage of using second-order information is that one can prove fast local convergence rates similar to the well-known results for the smooth Newton method. However, proximal (quasi-)Newton methods are in general only well-defined for convex $f$ and the convergence theorems typically require some strong convexity assumption.
In contrast, a proximal gradient method without second-order information can be shown to converge globally in the sense that every accumulation point of a sequence generated by this method is a stationary point of the objective function $\psi$, cf. Theorem 3.12, but it is not possible to achieve fast convergence results. The method presented in this chapter takes into account the advantages of both methods and combines them to get a globalized proximal Newton-type method. For this purpose, we use a novel descent condition to control which method is used in the current step.

We briefly discuss some related approaches on proximal Newton-type methods from the literature. In Lee, Sun, and Saunders [95] a generic version of the proximal Newton method is presented and several convergence results based on the exactness of the subproblem solutions and the Hessian approximation are stated. For the local convergence theory, they need strong convexity of $f$. In Yue, Zhou, and So [167], an inexact proximal Newton method with regularized Hessian is presented which assumes $f$ to be convex, but not strongly convex, and an error bound condition. Their inexactness criterion is similar to ours. The authors in [100, 152] assume that $f$ is convex and self-concordant and apply a damped proximal Newton method.
Bonettini et al. [25, 26] consider an inexact proximal gradient method with variable metric and an Armijo-type line search to solve problem (1.1). The structure of the method in [26] is similar to ours, but they have no globalization, have to add an overrelaxation step to ensure convergence and use a different inexactness criterion. Their convergence theory

covers global convergence and local convergence under the assumption that $\nabla f$ is Lipschitz continuous and $\psi$ satisfies the Kurdyka-Łojasiewicz property.
A similar method with various line search criteria is introduced by Lee and Wright [92]. Their inexactness criterion is related to the one from Bonettini et al. Furthermore, they use a backtracking strategy to update the matrix $H_k$ in (4.1), if suitable descent is not achieved. Here, convergence rates are proven for nonconvex as well as for convex problems.

The chapter is organized as follows. We first deduce the algorithmic framework and prove the well-definedness of our algorithm in Section 4.1. In Section 4.2 global convergence is proved under mild assumptions, whereas Section 4.3 deals with the proof of fast local convergence under the Kurdyka-Łojasiewicz property. The chapter is closed with some stronger local convergence results that hold for strongly convex functions in Section 4.4. Note that this chapter is mainly based on the author's work in [82].

## 4.1 Algorithmic Framework

We start with the derivation of our globalized inexact proximal Newton-type method. The main step of the method is the solution of the subproblem (1.2), where $H_k$ is an approximation to $\nabla^2 f(x^k)$. For the deduction of the inexact version, we consider smooth optimization problems for the moment. Here, one step of the classical version of Newton's method for minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$ consists in finding a search direction $d^k$ by solving $H_k d = -\nabla f(x^k)$. This is equivalent (assuming $H_k$ being positive definite for the moment) to solve the problem $\min_d f_k(d)$, where

$$f_k(d) := f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d \tag{4.2}$$

is a quadratic approximation of $f$ at the current iterate $x^k$. To solve this problem inexactly, one often uses the criterion

$$\|\nabla f_k(d)\| \leq \eta_k \|\nabla f(x^k)\| \tag{4.3}$$

for some $\eta_k \in (0,1)$, cf. [123, Section 7.1].

To adapt this strategy to the nonsmooth problem (1.1), we define

$$r_H(x) := \operatorname{prox}_\varphi^H \big(x - H^{-1}\nabla f(x)\big) - x = \underset{d\in\mathbb{R}^n}{\arg\min} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + \varphi(x+d) \right\}. \tag{4.4}$$

Note that actually this definition depends on the functions $f$ and $\varphi$, but as we view these as fixed within the current chapter, we omit this dependency in the notation. Furthermore, if $H = I$, we simply write $r(x) := r_I(x)$. If $\varphi \equiv 0$, i.e. in smooth optimization, note that $r(x)$ coincides with $-\nabla f(x)$. Moreover, we know that $r_H(x) = 0$ if and only if $x$ is a stationary point of $\psi$ (Proposition 3.7) and $r_H(x)$ is a continuous function of $x$ (Proposition 3.5). Due to these properties we use the residuum function $r(x)$ as a generalization of the derivative for the nonsmooth function $\psi = f + \varphi$.
Since there is no computationally inexpensive method to get a quadratic approximation to the nonsmooth function $\varphi$, we use the approximation

$$q_k(d) := f_k(d) + \varphi(x^k + d) = f(x^k) + \nabla f(x^k)^T d + \tfrac{1}{2} d^T H_k d + \varphi(x^k + d) \tag{4.5}$$

for $\psi$. Since $q_k$ is another function of the form (1.1), one can use the same idea as above to replace the derivative $\nabla f_k(x^k + d)$ by

$$\begin{aligned} r^k(x^k + d) &:= \operatorname{prox}_\varphi \big(x^k + d - \nabla f_k(d)\big) - x^k - d \\ &= \operatorname{prox}_\varphi \big(x^k + d - \nabla f(x^k) - H_k d\big) - x^k - d. \end{aligned}$$

This observation motivates to replace the inexactness criterion (4.3) by a condition like $\|r_k(x^k + d)\| \leq \eta_k \|r(x^k)\|$ for some $\eta_k \geq 0$, see [38, 95].
The criterion originates directly from the smooth Newton method and considers the distance of a point from being a solution to the subproblem in some sense, but the disadvantage is that a proximity operator needs to be computed to verify this condition. Hence, it is only applicable if the computation of the proximity operator is inexpensive or analytically possible. For that reason, some inexact proximal-type methods in the literature [26, 92] use a different inexactness criterion considering the value of the difference $q_k(x^k + d) - q_k(x^k + d^k_{ex})$ of the function values of the quadratic approximation $q_k$, where $d^k_{ex}$ is an exact minimizer of $q_k$. This type of inexactness criteria is also not ideal, since the solution $d^k_{ex}$ is of course unknown.

The main idea of our globalized proximal Newton-type method is similar to a standard globalization of the classical Newton method for smooth unconstrained optimization problems: Whenever the proximal Newton-type direction exists and satisfies a suitable sufficient decrease condition, the proximal Newton-type direction is accepted and followed by a line search. Otherwise, a proximal gradient step is taken which always exists and guarantees suitable global convergence properties. In contrast to the general theory for the proximal gradient method in Section 3.2 we restrict the proximal gradient steps in this method to the original purpose taking $H_k = \tau_k I$ for some $\tau_k > 0$, which avoids the need of using another algorithm for the subproblem in many cases of interest, where the proximity operator $\operatorname{prox}_\varphi$ can be computed analytically. The descent criterion used here is motivated by similar conditions in [53, 133]. The line search is based on the Armijo-type condition already used in the proximal gradient method and makes use of the same $\Delta_k$. The exact statement of our method is given in Algorithm 4.1, where, for the moment we allow $H_k$ to be an arbitrary symmetric matrix.
Before we start to analyse the convergence properties of Algorithm 4.1, let us add a few comments. The properties of Algorithm 4.1 obviously depend on the choice of the matrices $H_k$ and the degree of inexactness that is used to compute the inexact proximal Newton-type direction in (S.1). This degree is specified by the test in (4.6). Although parts of the local convergence theory require some additional assumptions regarding the choice of the sequence $\{\eta_k\}$, the global convergence analysis only depends on the choice $\eta_k \in [0, \eta)$ for some given $\eta \in (0, 1)$ and does not need the second condition in (4.6). The second condition in (4.6) is a safeguard which simplifies our local convergence theory. It guarantees the boundedness of the inexact proximal Newton direction which may not hold in general. In practice, this condition is very weak, and, in some situations, this second condition is not required explicitly, because it follows from other assumptions. Both conditions certainly hold for the exact solution of the corresponding subproblem, cf. Lemma 3.9 for the second condition and note that $\zeta < \frac{1}{2}$. Furthermore, the matrices $H_k$ do not need to be positive (semi-)definite for the global convergence analysis. In contrast, in the local convergence theory, these matrices have to be chosen to be appropriate approximations to the Hessian of $f$.
For that reason, the proximal subproblems in (S.1) of Algorithm 4.1 are not guaranteed to have a solution. The same difficulty arises within the classical Newton method since, in the indefinite case, the quadratic subproblem (4.2) certainly has no minimizer. Nevertheless,

**Algorithm 4.1** Globalized Inexact Proximal Newton-type Method (GPN)

(S.0) Choose $x^0 \in \operatorname{dom}\varphi$, initial parameters $\rho > 0$, $p > 2$, $\beta, \eta \in (0,1)$, $\sigma \in (0, \frac{1}{2})$, $\zeta \in (\sigma, \frac{1}{2})$, $0 < \tau_{\min} \le \tau_{\max}$, and set $k := 0$.

(S.1) Choose $H_k \in \mathbb{S}^n$, $\eta_k \in [0, \eta)$ and compute an inexact solution $d^k$ of the subproblem $\min_d q_k(d)$ with $q_k$ defined in (4.5) satisfying

$$\|r_k(x^k + d^k)\| \le \eta_k \|r(x^k)\| \qquad \text{and} \qquad q_k(d^k) - q_k(0) \le \zeta \Delta_k, \tag{4.6}$$

where $\Delta_k = \nabla f(x^k)^T d^k + \varphi(x^k + d^k) - \varphi(x^k)$. If this is not possible or the condition

$$\Delta_k \le -\rho \|d^k\|^p \tag{4.7}$$

is not satisfied, choose $\tau_k \in [\tau_{\min}, \tau_{\max}]$ and set $d^k = r_{\tau_k I}(x^k)$.

(S.2) If $d^k = 0$: STOP.

(S.3) Compute $t_k = \max\{\beta^l : l = 0, 1, 2, \dots\}$ such that

$$\psi(x^k + t_k d^k) \le \psi(x^k) + \sigma t_k \Delta_k. \tag{4.8}$$

(S.4) Set $x^{k+1} = x^k + t_k d^k$, update $k \leftarrow k + 1$ and go to (S.1).

the classical Newton method is often quite successful even if $H_k$ is indefinite (at least during some intermediate iterations), and the Newton direction is usually well-defined because it just computes a stationary point of the subproblem (4.2) which exists also for indefinite matrices $H_k$. Here, the situation is similar since the conditions (4.6) only check whether we have an (inexact) stationary point. Moreover, the circumstances here are even better than in the classical case since the additional function $\varphi$ may guarantee the existence of a minimum even for indefinite $H_k$, e.g. if $\varphi$ has compact support as this occurs when $\varphi$ is the indicator function of a bounded feasible set.
The constraint in (4.7) is a sufficient decrease condition, with $\rho > 0$ typically being a small constant.

For our subsequent analysis, we set

$$\begin{aligned}
\mathcal{K}_G : &= \{k \ge 0 : x^{k+1} \text{ was generated using a proximal gradient} \\
&\qquad\qquad\qquad\qquad \text{step satisfying } d^k = r_{\tau_k I}(x^k)\}, \\
\mathcal{K}_N : &= \{k \ge 0 : x^{k+1} \text{ was generated using an inexact proximal} \\
&\qquad\qquad\qquad\qquad \text{Newton-type step satisfying (4.6)}\}.
\end{aligned}$$

In addition, it has to be mentioned that the notation for the search direction is slightly inconsistent. Whenever we only use the term $d^k$, we refer $d^k$ obtained by the (inexact) proximal Newton step, if $k \in \mathcal{K}_N$, and $d^k = r_{\tau_k I}(x^k)$ obtained from a proximal gradient step for $k \in \mathcal{K}_G$. To emphasize that we refer to the search direction determined by the inexact proximal Newton method in (4.6) (regardless of whether it satisfies the descent criterion (4.7)), we use the term $d^k_N$. The same holds for $\Delta_k$.
The following result, which is [82, Proposition 3.2], shows that the step size rule in (S.3) is well-defined and Algorithm 4.1 is a descent method.

**Proposition 4.1.** *Consider a fixed iteration $k$ and suppose that $d^k \neq 0$. Then the line search in (S.3) is well-defined and yields a new iterate $x^{k+1}$ satisfying $\psi(x^{k+1}) < \psi(x^k)$.*

*Proof.* Since the proximal gradient method is well-defined by Proposition 3.10, the claim holds for $k \in \mathcal{K}_G$. Now, assume $k \in \mathcal{K}_N$, in which case (4.7) holds. Then $\Delta_k < 0$ and, therefore, the remaining part of the proof is identical to the one of Proposition 3.10. □

Proposition 4.1 requires $d^k \neq 0$. In view of the following result, this assumption can be stated without loss of generality. In particular, this result justifies our termination criterion in (S.2). Note that it coincides with [82, Lemma 3.3].

**Lemma 4.2.** *An iterate $x^k$ generated by Algorithm 4.1 is a stationary point of $\psi$ if and only if $d^k = 0$.*

*Proof.* For $k \in \mathcal{K}_G$, the result follows from Proposition 3.7. Hence assume $k \in \mathcal{K}_N$, and let $d^k = 0$. Since $r^k(x^k + d^k) = r(x^k)$, condition (4.6) yields $\|r(x^k)\| \leq \eta_k \cdot \|r(x^k)\|$. As $\eta_k \in [0, 1)$, we get $r(x^k) = 0$ and $x^k$ is a stationary point of $\psi$, using again Proposition 3.7. Conversely, assume that $d^k \neq 0$ for $k \in \mathcal{K}_N$. Then, analogous to Lemma 3.9, we get

$$\psi'(x^k; d^k) \leq \Delta_k \leq -\rho \|d^k\|^p < 0.$$

Hence $x^k$ is not a stationary point of $\psi$. □

Altogether, the previous results show that Algorithm 4.1 is well-defined.

## 4.2 Global Convergence Theory

In the following, we will prove global convergence results for the globalized inexact proximal Newton-type method in Algorithm 4.1. For this purpose, we assume that the method generates an infinite sequence $\{x^k\}$ such that $d^k \neq 0$ holds for all $k \geq 0$, which means by Lemma 4.2 that the sequence does not terminate with a stationary point of $\psi$ after finitely many iterations.

The following is the main global convergence result for Algorithm 4.1 from [82]. It guarantees stationarity of any accumulation point. In particular, if $\psi$ is convex, this implies that any accumulation point is a solution of the composite optimization problem (1.1).

**Theorem 4.3.** *Let $\{H_k\} \subset \mathbb{S}^n$ be a bounded sequence. Then every accumulation point of a sequence $\{x^k\}$ generated by Algorithm 4.1 is a stationary point of $\psi$.*

*Proof.* Let $\{x^k\}$ be a sequence generated by Algorithm 4.1 and $\mathcal{K} \subset \mathbb{N}_0$ such that $\{x^k\}_\mathcal{K}$ converges to some $x^* \in \mathbb{R}^n$. If there are infinitely many indices $k \in \mathcal{K}$ with $k \in \mathcal{K}_G$, i.e. the subsequence contains infinitely many iterates $x^k$ such that $x^{k+1}$ is generated by the proximal gradient method, Proposition 4.1 and the global convergence theorem for the proximal gradient method, Lemma 3.11, yield that $x^*$ is a stationary point of $\psi$.

Hence, consider the case $\mathcal{K} \subset \mathcal{K}_N$, where all elements of the subsequence $\{x^{k+1}\}_\mathcal{K}$ are generated by inexact proximal Newton-type steps. In that case the proof follows the same ideas as the one of Lemma 3.11. Since $\{\psi(x^k)\}$ is monotonically decreasing by Proposition 4.1, $\{x^k\}_\mathcal{K}$ converges to $x^*$, and since $\psi$ is lower semicontinuous, we get the convergence of the entire sequence $\{\psi(x^k)\}$ to some finite number $\psi^*$. The Armijo-type line search (4.8) therefore yields

$$0 \leftarrow \psi(x^{k+1}) - \psi(x^k) \leq \sigma t_k \Delta_k < 0$$

and, hence, $t_k \Delta_k \to 0$ for $k \to \infty$. We claim that this implies $\{\|d^k\|\}_\mathcal{K} \to 0$ (possibly after taking another subsequence). To verify this statement, we distinguish two cases:

*Case 1:* $\liminf_{k\in\mathcal{K}} t_k > 0$. Then $\{\Delta_k\}_{\mathcal{K}} \to 0$, and we obtain $\{\|d^k\|\}_{\mathcal{K}} \to 0$ in view of (4.7).
*Case 2:* $\liminf_{k\in\mathcal{K}} t_k = 0$. Without loss of generality, assume $\lim_{k\in\mathcal{K}} t_k = 0$. Then, for all $k \in \mathcal{K}$ sufficiently large, there holds $t_k \le \beta$ and the line search test is violated for the step size $\hat{t}_k := t_k/\beta$. Using the monotonicity of the difference quotient of convex functions, cf. Proposition 2.28(a), and the definition of $\Delta_k$, we therefore obtain

$$\begin{aligned}
\sigma\Delta_k &< \frac{\psi(x^k + \hat{t}_k d^k) - \psi(x^k)}{\hat{t}_k} \\
&\le \frac{f(x^k + \hat{t}_k d^k) - f(x^k)}{\hat{t}_k} + \varphi(x^k + d^k) - \varphi(x^k) \\
&= \frac{f(x^k + \hat{t}_k d^k) - f(x^k)}{\hat{t}_k} - \nabla f(x^k)^T d^k + \Delta_k \\
&= \big(\nabla f(\xi^k) - \nabla f(x^k)\big)^T d^k + \Delta_k
\end{aligned}$$

for all sufficiently large $k \in \mathcal{K}$, where the last expression uses the mean value theorem with some $\xi^k \in (x^k, x^k + \hat{t}_k d^k)$. Reordering these expressions, we get

$$0 < -(1-\sigma)\Delta_k < \big(\nabla f(\xi^k) - \nabla f(x^k)\big)^T d^k. \tag{4.9}$$

Using (4.7) yields

$$(1-\sigma)\rho\|d^k\|^{p-1} \le \|\nabla f(\xi^k) - \nabla f(x^k)\| \tag{4.10}$$

for all $k \in \mathcal{K}$. By the assumption of case 2 we have $\{t_k\Delta_k\}_{\mathcal{K}} \to 0$. In view of (4.7), this yields $\{t_k\|d^k\|^p\}_{\mathcal{K}} \to 0$. Since $p > 1$, this implies $\{\hat{t}_k\|d^k\|\}_{\mathcal{K}} \to 0$. Hence, the right hand side of (4.10) converges to zero due to the uniform continuity of $\nabla f$ on compact sets. Consequently, (4.10) shows that $\|d^k\| \to_{\mathcal{K}} 0$, noting again that $p > 1$.
Therefore, $d^k \to_{\mathcal{K}} 0$ holds in both cases. Since $x^k \to_{\mathcal{K}} x^*$, the definition of $d^k$ also implies $x^k + d^k \to_{\mathcal{K}} x^*$. Using the continuity of the proximity operator, we therefore get

$$r(x^k) \to_{\mathcal{K}} \operatorname{prox}_\varphi\big(x^* - \nabla f(x^*)\big) - x^*$$

and, since $\{H_k\}$ is bounded by assumption, the definition of $r^k$ yields

$$r^k(x^k + d^k) \to_{\mathcal{K}} \operatorname{prox}_\varphi\big(x^* - \nabla f(x^*)\big) - x^*.$$

Since $\|r^k(x^k + d^k)\| \le \eta\|r(x^k)\|$ for all $k \in \mathcal{K}$ in view of (4.6) and $\eta \in (0,1)$, taking the limit $k \in \mathcal{K}$, $k \to \infty$ therefore implies $\lim_{k\in\mathcal{K},k\to\infty} r(x^k) = r(x^*) = 0$ and therefore $x^* = \operatorname{prox}_\varphi\big(x^* - \nabla f(x^*)\big)$, which is equivalent to $x^*$ being a stationary point of $\psi$, cf. Corollary 3.8. □

Note that the proof of Theorem 4.3 only requires $p > 1$ and the first condition from (4.6). The second condition from (4.6) is only needed in the local convergence theory, whereas $p > 2$ is already a preliminary of the next result, which gives more information about the sequences $\{\Delta_k\}$ and $\{d^k\}$ under common assumptions.

**Theorem 4.4.** *Let $\nabla f$ be Lipschitz continuous and $\{x^k\}$ a sequence generated by Algorithm 4.1 such that $\{\psi(x^k)\}$ is bounded from below. Then the corresponding sequences $\{\Delta_k\}$ and $\{d^k\}$ converge to $0$. If, in addition, there exists an isolated accumulation point $x^*$ of $\{x^k\}$, the complete sequence converges to $x^*$.*

*Proof.* The arguments of the proof of Theorem 3.14 yield $\{\Delta_k\}_{\mathcal{K}_G} \to 0$ and $\{d^k\}_{\mathcal{K}_G} \to 0$. We combine the technique of that proof with some arguments from the proof of Theorem 4.3 to obtain the same result for the subsequences of elements in $\mathcal{K}_N$. For that purpose, note that since $\{\psi(x^k)\}$ is bounded from below and monotonically decreasing (Proposition 4.1), the sequence converges to some $\psi^* \in \mathbb{R}$. Hence, summation of the Armijo-type condition (4.8) yields

$$\psi(x^0) - \psi^* = \sum_{k=0}^{\infty} \psi(x^k) - \psi(x^{k+1}) \geq \sum_{k \in \mathcal{K}_N} \psi(x^k) - \psi(x^{k+1}) \geq \sum_{k \in \mathcal{K}_N} -\sigma t_k \Delta_k \geq 0,$$

where we used that $\Delta_k$ is nonpositive for all $k \geq 0$. Hence, $\{t_k \Delta_k\}_{\mathcal{K}_N} \to 0$. Assume that $\{\Delta_k\}_{\mathcal{K}_N} \not\to 0$. Then there exists $\mathcal{K} \subset \mathcal{K}_N$ and $\delta > 0$ such that $\Delta_k \leq -\delta$ for all $k \in \mathcal{K}$ (note, again, that $\Delta_k$ is nonpositive), and we have $\{t_k\}_{\mathcal{K}} \to 0$. Similar to (4.9) in the proof of Theorem 4.3 we get

$$(1-\sigma)\delta \leq -(1-\sigma)\Delta_k < \left(\nabla f(\xi^k) - \nabla f(x^k)\right)^T d^k \leq L(t_k/\beta)\|d^k\|^2 \tag{4.11}$$

for some $\xi^k \in (x^k, x^k + t_k/\beta\ d^k)$, where we used the Lipschitz continuity of $\nabla f$ with Lipschitz constant $L > 0$. Further, the arguments from the proof of Theorem 4.3 yield $\{t_k\|d^k\|^2\}_{\mathcal{K}} \to 0$ (note, that we need $p > 2$ here). Hence, the right hand side of (4.11) converges to 0, which is a contradiction. Thus, $\{\Delta_k\}_{\mathcal{K}_N} \to 0$ and (4.7) yield $\{d^k\}_{\mathcal{K}_N} \to 0$. The remaining part of the proof is exactly as in the proof of Theorem 3.14. □

We note that the assumption on $\{\psi(x^k)\}$ in Theorem 4.4 is satisfied whenever the sequence $\{x^k\}$ has an accumulation point or the function $\psi$ itself is bounded from below.

## 4.3 Local Convergence Theory for KL-Functions

We now turn to the local convergence properties of Algorithm 4.1. In recent years, increasing importance has been given to convergence theory under the Kurdyka-Łojasiewicz property, cf. Section 2.3. This property holds for strongly convex functions, but also for numerous further examples relevant in application.

For that reason, the theory in this section is based on the assumption that $\psi$ is a Kurdyka-Łojasiewicz function and $x^* \in \mathbb{R}^n$ is an isolated stationary point of $\psi$. Furthermore, we still assume that the method generates an infinite sequence $\{x^k\}$ such that $d^k \neq 0$ holds for all $k \geq 0$. Under these assumptions, if $x^*$ is an accumulation point of the sequence $\{x^k\}$, by Theorem 4.4 the complete sequence converges to $x^*$ (note that it is enough to assume local Lipschitz continuity in a neighbourhood of $x^*$).

For the following analysis, we assume in addition, that the sequence $\{H_k\}$ is uniformly bounded and positive definite, i.e. there exist $0 < m \leq M$ such that $mI \preceq H_k \preceq MI$ holds for all $k \geq 0$. Under suitable further preliminaries we prove that the method finally performs only iterates with $k \in \mathcal{K}_N$ and always the full step length $t_k = 1$ is attained. The main steps into this direction are summarized in the following observations, which are essentially parts of Lemma 4.4 and Theorem 4.5 in [82].

**Proposition 4.5.** *Consider Algorithm 4.1 with $\{H_k\}$ satisfying $mI \preceq H_k \preceq MI$ for all $k \geq 0$ with suitable $0 < m \leq M$, and let $x^*$ be a stationary point of $\psi$. Then there exist constants $\varepsilon, C, \kappa > 0$ such that, for any iterate $x^k \in B_\varepsilon(x^*)$, the following statements hold, where $d^k_{ex}$ is the exact solution of the corresponding subproblem in (S.1) of Algorithm 4.1:*

*(a)* $\left\|d^k - d^k_{ex}\right\| \leq C\eta_k\|r(x^k)\|$.

*(b)* $\left\|d^k_{ex}\right\| \leq \kappa\|x^k - x^*\|$.

*Proof.* We verify the statements separately, using possibly different values of $\varepsilon$.
(a) First, note that the function $q_k$ in (4.5) is strongly convex and, therefore, has a unique minimizer (Corollary 2.17). Thus, the exact solution

$$d_{ex}^k = r_{H_k}(x^k) = \operatorname{prox}_\varphi^{H_k}\big(x^k - H_k^{-1}\nabla f(x^k)\big) - x^k$$

of the subproblem exists and hence guarantees that there is a possibly inexact solution $d^k$. Furthermore, set $y^k = x^k + d^k$.
Since $r^k(y^k) = \operatorname{prox}_\varphi\big(y^k - \nabla f_k(y^k)\big) - y^k$, we obtain from Lemma 3.4 that

$$-\nabla f_k(y^k) - r^k(y^k) \in \partial\varphi(y^k + r^k(y^k)).$$

The definition of $q_k$ together with the subdifferential sum rule (Proposition 2.45) therefore implies

$$-r^k(y^k) + \nabla f_k(y^k + r^k(y^k)) - \nabla f(y^k) \in \partial q_k\big(d^k + r^k(y^k)\big),$$

which is equivalent to

$$(H_k - I)r^k(y^k) \in \partial q_k\big(d^k + r^k(y^k)\big). \tag{4.12}$$

Since $q_k$ is strongly convex with modulus $m > 0$, its subdifferential is strongly monotone in this neighbourhood with the same modulus (Proposition 2.22). Hence, using (4.12) together with $0 \in \partial q_k(d_{ex}^k)$, cf. Proposition 2.47, we get

$$\big\langle (H_k - I)r^k(y^k), d^k + r^k(y^k) - d_{ex}^k\big\rangle \geq m\big\|d^k + r^k(y^k) - d_{ex}^k\big\|^2.$$

Applying the Cauchy-Schwarz inequality, this implies

$$\big\|d^k + r^k(y^k) - d_{ex}^k\big\| \leq \frac{1}{m}\big\|(H_k - I)r^k(y^k)\big\| \leq \frac{1}{m}(1+M)\|r^k(y^k)\|.$$

Using the inexactness criterion (4.6), we finally get

$$\begin{aligned}\|d^k - d_{ex}^k\| &\leq \|d^k + r^k(y^k) - d_{ex}^k\| + \|r^k(y^k)\| \\ &\leq \frac{1}{m}(1+M)\|r^k(y^k)\| + \|r^k(y^k)\| \leq C\eta_k\|r(x^k)\|\end{aligned}$$

with $C := (1 + M + m)/m$.
(b) Using Lemma 3.6 and the uniform boundedness of $\{H_k\}$, there exists $\hat\kappa > 0$ such that $\|d_{ex}^k\| = \|r_{H_k}(x^k)\| \leq \hat\kappa\|r(x^k)\|$. Let $\varepsilon > 0$ be such that $\nabla f$ is Lipschitz continuous with Lipschitz constant $L > 0$ in $B_\varepsilon(x^*)$. Thus, using $r(x^*) = 0$, cf. Proposition 3.7 and the definition of the mapping $r$, and the nonexpansivity of the proximity operator (Proposition 3.5), we get

$$\begin{aligned}\|d_{ex}^k\| &\leq \hat\kappa\|r(x^k)\| = \hat\kappa\|r(x^k) - r(x^*)\| \\ &= \hat\kappa\big\|\operatorname{prox}_\varphi(x^k - \nabla f(x^k)) - x^k - \operatorname{prox}_\varphi(x^* - \nabla f(x^*)) + x^*\big\| \\ &\leq \hat\kappa\big(\big\|\operatorname{prox}_\varphi(x^k - \nabla f(x^k)) - \operatorname{prox}_\varphi(x^* - \nabla f(x^*))\big\| + \|x^k - x^*\|\big) \\ &\leq \hat\kappa\big(\|\nabla f(x^k) - \nabla f(x^*)\| + 2\|x^k - x^*\|\big) \leq \kappa\|x^k - x^*\|,\end{aligned}$$

with $\kappa = \hat\kappa(2 + L)$. □

For the following analysis, we assume in addition that $f$ is twice continuously differentiable in a neighbourhood of $x^*$, and the sequence $\{H_k\}$ satisfies the Dennis-Moré condition [56]

$$\lim_{k\to\infty} \frac{\left\|\left(H_k - \nabla^2 f(x^*)\right)d^k\right\|}{\|d^k\|} = 0. \tag{4.13}$$

Introduced to prove local convergence for smooth quasi-Newton methods, it is also predestined for the same purpose when considering proximal quasi-Newton methods, cf. [95].

**Remark 4.6.** Under appropriate assumptions, it can be shown that the sequence $\{H_k\}$ used in Algorithm 4.1 satisfies the Dennis-Moré condition.
In detail, assume that $\{H_k\}$ is updated using the BFGS-formula mentioned in (3.15) and $\nabla^2 f$ is Lipschitz continuous in a neighbourhood of $x^*$. Furthermore, let the sequence $\{x^k\}$ have finite length, i.e.

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| < +\infty,$$

which was initially associated with quasi-Newton methods by Dennis and Moré [55]. This property can be proved in a similar way to [26, Theorem 1] under the preliminaries of Theorem 4.8. Then, the structure of the proof of this claim follows the one in [37], see also Theorem 6.6 in [123] for smooth quasi-Newton methods and Lemma 3 in the appendix of [169]. $\diamondsuit$

A suitable combination of the previous results leads to the following global and local convergence result for Algorithm 4.1.

**Theorem 4.7.** *Consider Algorithm 4.1 with $\{H_k\}$ satisfying the Dennis-Moré condition and $mI \preceq H_k \preceq MI$ for all $k \geq 0$ with suitable $0 < m \leq M$. Let $x^*$ be an accumulation point of a sequence $\{x^k\}$ generated by Algorithm 4.1, which is an isolated stationary point of $\psi$. Then the following statements hold:*

*(a) The entire sequence $\{x^k\}$ converges to $x^*$.*

*(b) For all sufficiently large $k \geq 0$, the search direction $d^k$ is attained by the inexact proximal Newton-type direction, i.e. $k \in \mathcal{K}_N$.*

*(c) For all sufficiently large $k \geq 0$, the full step length $t_k = 1$ is accepted.*

*Proof.* (a) Let $\mathcal{K} \subset \mathbb{N}$ be such that $\{x^k\}_{\mathcal{K}}$ converges to $x^*$. Using $mI \preceq H_k \preceq MI$, Lemma 3.6 and the continuity of the proximity operator, we get $\{d^k\}_{\mathcal{K}} \to 0$. Thus, Proposition 2.60 yields the claim.
(b) Similar to the proof of Proposition 4.5, there exists an inexact solution $d_N^k$ of the subproblem defined in (4.6) for all $k \geq 0$. We prove $d_N^k = d^k$ for sufficiently large $k \geq 0$, which follows, if the sufficient decrease condition (4.7) holds. For that purpose, let $\Delta_{k,N}$ be the $\Delta$-function corresponding to the search direction $d_N^k$, i.e. $\Delta_{k,N} := \nabla f(x^k)^T d_N^k + \varphi(x^k + d_N^k) - \varphi(x^k)$. Then the second condition in (4.6) is equivalent to

$$(1-\zeta)\Delta_{k,N} \leq -\frac{1}{2}(d_N^k)^T H_k d_N^k,$$

which yields

$$\Delta_{k,N} \leq -\tilde{c}\|d_N^k\|^2 \qquad \text{for } \tilde{c} := \frac{m}{2(1-\zeta)}. \tag{4.14}$$

Since $x^*$ is a stationary point of $\psi$, hence $r(x^*) = 0$, it follows from the continuity of $r$ and the results in Proposition 4.5 that the estimates

$$\|d_N^k - d_{ex}^k\| \leq \frac{1}{2}\left(\frac{\rho}{\tilde{c}}\right)^{1/(2-p)} \qquad \text{and} \qquad \|d_{ex}^k\| \leq \frac{1}{2}\left(\frac{\rho}{\tilde{c}}\right)^{1/(2-p)}$$

hold for all sufficiently large $k \geq 0$, where, again $d^k_{ex} = r_{H_k}(x^k)$ denotes the exact minimizer of $q_k$. Combining these inequalities yields $\|d^k_N\| \leq (\rho/\tilde{c})^{1/(2-p)}$. We therefore get

$$\Delta_{k,N} \leq -\tilde{c}\|d^k_N\|^2 = -\tilde{c}\|d^k_N\|^p\|d^k_N\|^{2-p} \leq -\rho\|d^k_N\|^p,$$

noting that $p > 2$. Thus, the sufficient descent condition (4.7) is fulfilled and the search direction $d^k = d^k_N$ is obtained by the inexact proximal Newton-type method.
(c) Taylor expansion yields

$$f(x^k + d^k) - f(x^k) = \nabla f(x^k)^T d^k + \frac{1}{2}(d^k)^T\nabla^2 f(x^k)d^k + \frac{1}{2}(d^k)^T\big(\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\big)d^k$$

for some $\xi^k \in (x^k, x^k + d^k)$. Hence, we get

$$\begin{aligned}
&\psi(x^k + d^k) - \psi(x^k) + q_k(0) - q_k(d^k)\\
&\qquad = f(x^k + d^k) - f(x^k) - \nabla f(x^k)^T d^k - \frac{1}{2}(d^k)^T H_k d^k\\
&\qquad \leq \frac{1}{2}\big\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\big\| \cdot \|d^k\|^2 + \frac{1}{2}\big\|\nabla^2 f(x^k) - \nabla^2 f(x^*)\big\| \cdot \|d^k\|^2\\
&\qquad\quad + \frac{1}{2}\big\|\big(H_k - \nabla^2 f(x^*)\big)d^k\big\| \cdot \|d^k\|.
\end{aligned}$$

Since $x^k \to x^*$ and $d^k \to 0$, the first and second term are of order $o(\|d^k\|^2)$ for $k \to \infty$. Furthermore, by the Dennis-Moré criterion, the same holds for the third term. As before, it follows from the continuity of $r$ and the results in Proposition 4.5 that $\|d^k\| \to 0$. Therefore, the above term is bounded by $(\zeta - \sigma)\tilde{c}\|d^k\|^2$ for all $k \geq 0$ sufficiently large. Thus, using (4.6), we obtain

$$\begin{aligned}
\psi(x^k + d^k) - \psi(x^k) &= \Big(\psi(x^k + d^k) - \psi(x^k) + q_k(0) - q_k(d^k)\Big) + q_k(d^k) - q_k(0)\\
&\leq (\zeta - \sigma)\tilde{c}\|d^k\|^2 + \zeta\Delta_k\\
&= (\zeta - \sigma)\tilde{c}\|d^k\|^2 + \sigma\Delta_k + (\zeta - \sigma)\Delta_k\\
&\leq (\zeta - \sigma)\tilde{c}\|d^k\|^2 + \sigma\Delta_k - (\zeta - \sigma)\tilde{c}\|d^k\|^2 = \sigma\Delta_k,
\end{aligned}$$

for all sufficiently large $k \geq 0$, where the final inequality follows from (4.7) (note that $\Delta_k = \Delta_{k,N}$ in the current situation). This proves that (4.8) holds with $t_k = 1$ and, hence, the full step length is attained. □

It remains to provide convergence rates for Algorithm 4.1. To this end, we finally need the assumption of the Kurdyka-Łojasiewicz-property in $x^*$ from Definition 2.52. Using this property we state the following result.

**Theorem 4.8.** *In addition to the assumptions of Theorem 4.7 assume that the sequence $\{x^k\}$ satisfies the following condition: For every $k \geq 0$ there exists $s^k \in \partial\psi(x^k + d^k)$ such that*

$$\|s^k\| \leq \alpha_1\|d^k\| + \alpha_2|\Delta_k| \tag{4.15}$$

*holds for some $\alpha_1, \alpha_2 \geq 0$ and sufficiently large $k \geq 0$. Furthermore, assume that the KL-property is satisfied in $x^*$ with the function $\phi(s) = \frac{C}{\theta}s^\theta$ for some $C > 0$ and $\theta \in (0, 1]$. Then the following hold:*

*(a) If $\theta = 1$, the method terminates after finitely many steps with the exact solution $x^*$.*

*(b) If $\theta \in [\frac{1}{2}, 1)$, there exists $\delta > 0$ such that*

$$\psi(x^k) - \psi(x^*) = \mathcal{O}(e^{-\delta k}), \quad \text{and} \quad \|x^k - x^*\| = \mathcal{O}(e^{-\delta/2 \cdot k}),$$

*i.e. both sequences converge R-linearly.*

*(c) If $\theta \in (0, \frac{1}{2})$, there exists $k_0 \geq 0$ such that*

$$\psi(x^k) - \psi(x^*) = \mathcal{O}\big((k - k_0)^{-\frac{1}{1-2\theta}}\big) \quad \text{and} \quad \|x^k - x^*\| = \mathcal{O}\big((k - k_0 + 1)^{-\frac{1}{1-2\theta}}\big).$$

The result is similar to [26, Theorem 3], but there the authors need an overrelaxation step to guarantee convergence under the KL-property. In detail, the update in (S.4) of Algorithm 4.1 is replaced by

$$x^{k+1} = \begin{cases} x^k + d^k, & \text{if } \psi(x^k + d^k) \leq \psi(x^k + t_k d^k), \\ x^k + t_k d^k, & \text{otherwise.} \end{cases}$$

Since we have seen in Theorem 4.7 that our method finally makes only full steps, this overrelaxation is not necessary and we can mainly apply the convergence theory of [26]. The main difference is that we need to adjust their setting to our inexactness criterion (4.6). Hence, the proofs are very similar. We therefore skip them and refer to the analysis in [26]. For the sake of completeness, however, details of the proof are also provided in the appendix.

It seems necessary to make some comments regarding the assumption (4.15). First, consider the case that the subproblems in Algorithm 4.1 are solved exactly, hence $d^k$ minimizes $q_k$. From Fermat's rule (Proposition 2.47), we get

$$0 \in \nabla f(x^k) + H_k d^k + \partial\varphi(x^k + d^k),$$

which is equivalent to

$$s^k := \nabla f(x^k + d^k) - \nabla f(x^k) - H_k d^k \in \partial\psi(x^k + d^k).$$

Hence, if $x^k$ is sufficiently close to $x^*$ such that $\nabla f$ is Lipschitz continuous with Lipschitz constant $L > 0$ in an appropriate neighbourhood, we have $\|s^k\| \leq (L + M)\|d^k\|$. Thus, (4.15) holds with $\alpha_1 = L + M$ and $\alpha_2 = 0$. Another motivation, which shows that this assumption is reasonable also for the inexact solution, is provided with the convergence proof in the appendix.

**Remark 4.9.** Note that for the convergence rates in Theorem 4.8 we do not assume that $\eta$ has some sufficiently small value or the sequence $\{\eta_k\}$ converges to 0. Instead we get the same convergence rates as proven for the exact solution of the subproblems. Nevertheless, the deduction of the results in the appendix show that the constant, which is implicitly stated in the $\mathcal{O}$-notation, depends on $\eta$. Thus, if $\eta_k \to 0$, in the author's opinion it should be possible to arrive at an even stronger result, replacing $\mathcal{O}$ by $o$, which may be part of some future research. ◊

## 4.4 Local Convergence under Strong Convexity

Although the above local convergence theory already covers strongly convex functions $\psi$ (Corollary 2.56), some of the statements can be simplified and stronger convergence results can be derived. Therefore, we use the previous results in combination with the approach in

[82] to consider this case separately in the following. Note that the results in this section are in parts taken from [82].
For the purpose of our analysis, we assume that $\psi$ is locally strongly convex in a neighbourhood of an accumulation point $x^*$ of a sequence of iterates generated by Algorithm 4.1. Note that this assumption certainly holds if the Hessian $\nabla^2 f(x^*)$ is positive definite. Furthermore, as before, we assume that the sequence $\{H_k\}$ is uniformly bounded and positive definite and satisfies the Dennis-Moré-condition (4.13). Under these preliminaries we first show that in this case the assumptions can be weakened, as we do not need to assume that $x^*$ is an isolated accumulation point explicitly.

**Proposition 4.10.** *Let $x^*$ be an accumulation point of a sequence $\{x^k\}$ generated by Algorithm 4.1 such that $\psi$ is strongly convex in a neighbourhood of $x^*$. Then the complete sequence $\{x^k\}$ converges to $x^*$, and $x^*$ is a strict local minimizer of $\psi$.*

*Proof.* In view of Theorem 4.3, every accumulation point of the sequence $\{x^k\}$ is a stationary point of $\psi$. Since $\psi$ is locally strongly convex, $x^*$ is the only stationary point in a suitable neighbourhood. Hence, $x^*$ is necessarily the only accumulation point of the sequence $\{x^k\}$ in this neighbourhood, and a strict local minimum of $\psi$. Since $\{\psi(x^k)\}$ is bounded from below by $\psi(x^*)$ and $\nabla f$ is locally Lipschitz continuous in a neighbourhood of $x^*$, we apply Theorem 4.4, which directly yields the convergence of the complete sequence to $x^*$. □

For the convergence result, we need one more technical estimate, which is stated in the following.

**Lemma 4.11.** *Consider Algorithm 4.1 with $\{H_k\}$ satisfying the Dennis-Moré condition* (4.13) *and $mI \preceq H_k \preceq MI$ for all $k \geq 0$ with suitable $0 < m \leq M$. Let $x^*$ be a stationary point of $\psi$ such that $\psi$ is locally strongly convex in a neighbourhood of $x^*$. Then there exist constants $\varepsilon, C', \mu > 0$ such that, for any iterate $x^k \in B_\varepsilon(x^*)$, there holds*

$$\begin{aligned}\mu\|x^k + d^k_{ex} - x^*\| \leq & C'\eta_k\|r(x^k)\| + \big\|\big(H_k - \nabla^2 f(x^*)\big)d^k\big\| \\ & + \big\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\big\|,\end{aligned}$$

*where $d^k_{ex}$ denotes the exact solution of the subproblem $\min_d q_k(d)$.*

*Proof.* As in the proof of Proposition 4.5 note that the function $q_k$ is strongly convex and, therefore, has a unique minimizer. Thus, the exact solution $d^k_{ex} = r_{H_k}(x^k)$ of the subproblem exists and, hence, also guarantees the existence of the (possibly inexact) solution $d^k$.
Set $y^k_{ex} := x^k + d^k_{ex}$ and note that the above inequality holds trivially for $y^k_{ex} = x^*$. Thus, assume $y^k_{ex} \neq x^*$. First, note that Proposition 4.5(a) implies

$$\begin{aligned}\|(H_k - \nabla^2 f(x^*))d^k_{ex}\| &\leq (M + \|\nabla^2 f(x^*)\|)\|d^k_{ex} - d^k\| + \|(H_k - \nabla^2 f(x^*))d^k\| \\ &\leq C(M + \|\nabla^2 f(x^*)\|)\eta_k\|r(x^k)\| + \|(H_k - \nabla^2 f(x^*))d^k\|. \qquad (4.16)\end{aligned}$$

Since $\psi$ is locally strongly convex in a neighbourhood of $x^*$, its subdifferential is strongly monotone (Proposition 2.22), i.e. there exist $\varepsilon > 0$ and $\mu > 0$ such that

$$\langle x - y, \nabla f(x) + s(x) - \nabla f(y) - s(y)\rangle \geq 2\mu\|x - y\|^2$$

holds for all $x, y \in B_\varepsilon(x^*)$ and $s(x) \in \partial\varphi(x), s(y) \in \partial\varphi(y)$. Using the stationarity of $x^*$ and $y^k_{ex}$, we have $0 \in \nabla f(x^*) + \partial\varphi(x^*)$ and $0 \in \nabla f(x^k) + H_k d^k + \partial\varphi(y^k_{ex})$ by Fermat's rule (Proposition 2.47). Thus, also noting that $y^k_{ex}$ eventually belongs to $B_\varepsilon(x^*)$ in view of

Proposition 4.5(b), we get

$$\begin{aligned}2\mu\|y_{ex}^k - x^*\|^2 \leq & \langle \nabla f(y_{ex}^k) - \nabla f(x^k) - H_k d_{ex}^k, y_{ex}^k - x^* \rangle \\ = & \langle \big(\nabla^2 f(x^*) - H_k\big) d_{ex}^k, y_{ex}^k - x^* \rangle \\ & + \langle \nabla f(x^k) - \nabla f(y_{ex}^k) + \nabla^2 f(x^*) d_{ex}^k, y_{ex}^k - x^* \rangle \\ \leq & \big\| \big(\nabla^2 f(x^*) - H_k\big) d_{ex}^k \big\| \cdot \|y_{ex}^k - x^*\| \\ & + \|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\| \cdot \|y_{ex}^k - x^*\| \\ & + \|\nabla f(x^*) - \nabla f(y_{ex}^k) - \nabla^2 f(x^*)(x^* - y_{ex}^k)\| \cdot \|x^* - y_{ex}^k\|.\end{aligned}$$

Using Proposition 4.5(b) again and reducing $\varepsilon > 0$, if necessary, we get

$$\|\nabla f(x^*) - \nabla f(y_{ex}^k) - \nabla^2 f(x^*)(x^* - y_{ex}^k)\| \leq \mu \|x^* - y_{ex}^k\|$$

from the twice continuous differentiability of $f$. The assertion follows from dividing by $\|x^* - y^k\|$ and using (4.16). □

Furthermore, we get the following interesting relation between $\|r(x)\|$ and the distance of $x$ to the optimal solution $x^*$:

**Proposition 4.12.** *Let $x^*$ be an isolated stationary point of $\psi$ and $\varepsilon > 0$. Then the following hold:*

*(a) If $\nabla f$ is Lipschitz continuous in $B_\varepsilon(x^*)$ with Lipschitz constant $L > 0$, then*

$$\|r(x)\| \leq (2 + L)\|x - x^*\|$$

*for all $x \in B_\varepsilon(x^*)$.*

*(b) If, in addition, $\psi$ is strongly convex with modulus $\mu > 0$, then*

$$\|x - x^*\| \leq (1 + L)(1 + \tfrac{1}{\mu})\|r(x)\|$$

*for all $x \in B_\varepsilon(x^*)$.*

*Proof.* (a) Since $x^*$ is a stationary point of $\psi$, we know that $r(x^*) = 0$ (Proposition 3.7 and the definition of $r$). Using the nonexpansivity of the proximity operator (Proposition 3.5) and the Lipschitz continuity of $\nabla f$, we get

$$\begin{aligned}\|r(x)\| &= \|r(x) - r(x^*)\| \\ &= \|x - \operatorname{prox}_\varphi(x - \nabla f(x)) - x^* + \operatorname{prox}_\varphi(x^* - \nabla f(x^*))\| \\ &\leq 2\|x - x^*\| + \|\nabla f(x) - \nabla f(x^*)\| \leq (2 + L)\|x - x^*\|.\end{aligned}$$

(b) The result holds trivially for $x = x^*$, hence assume that $x \neq x^*$. Using the strong convexity of $\psi$, we get $\psi(x) - \psi(x^*) \geq \mu\|x - x^*\|^2$. Let $v \in \partial\psi(x)$. Then, the definition of the convex subdifferential yields $v^T(x - x^*) \geq \psi(x) - \psi(x^*)$. Combining both estimates results in

$$\|v\| \geq \frac{\psi(x) - \psi(x^*)}{\|x - x^*\|} \geq \mu\|x - x^*\|.$$

Hence, we get $\|x - x^*\| \leq \frac{1}{\mu}\operatorname{dist}(0, \partial\psi(x))$. The claim follows from [114, Proposition 2.1], see also Theorems 3.4 and 3.5 in [58] for the proof. □

A combination of the previous results leads to the following convergence result for Algorithm 4.1 under the main assumption of local strong convexity.

**Theorem 4.13.** *Consider Algorithm 4.1 and assume that the sequence $\{H_k\}$ satisfies the assumptions from Lemma 4.11. Let $x^*$ be an accumulation point of a sequence $\{x^k\}$ generated by Algorithm 4.1 such that $\psi$ is locally strongly convex in a neighbourhood of $x^*$. Then the following statements hold:*

*(a) For all sufficiently large $k \geq 0$, the search direction is attained by the inexact proximal Newton-type direction, i.e. $k \in \mathcal{K}_N$.*

*(b) For all sufficiently large $k \geq 0$, the full step size $t_k = 1$ is accepted.*

*(c) If $\eta < \overline{\eta}$, the sequence $\{x^k\}$ converges linearly to $x^*$, where*

$$\overline{\eta} = 1/((C + \frac{1}{\mu}C')(L+2))$$

*with $C, C', \mu$ from Proposition 4.5 and Lemma 4.11, and a local Lipschitz constant $L > 0$ of $\nabla f$ in a neighbourhood of $x^*$.*

*(d) If $\{\eta_k\} \to 0$, the sequence $\{x^k\}$ converges superlinearly to $x^*$.*

*Proof.* Note that by Proposition 4.10 $x^*$ is both a stationary point and a strict local minimum of $\psi$, and that the entire sequence $\{x^k\}$ converges to $x^*$. Part (a) and (b) coincide with the corresponding statements in Theorem 4.7.

For the remaining part choose $\varepsilon > 0$ such that Proposition 4.5 and Lemma 4.11 hold for $x^k \in B_\varepsilon(x^*)$ and $\nabla f$ is Lipschitz continuous with constant $L > 0$ in $B_\varepsilon(x^*)$. Let $k_0 \geq 0$ be sufficiently large such that all iterates $x^k$ for $k \geq k_0$ are in this neighbourhood and (a) and (b) hold for these iterates. Using parts (a) and (b) yields $x^{k+1} = x^k + d^k$, where $d^k$ is an inexact proximal Newton-type step. Thus, by Proposition 4.5(a) and Lemma 4.11, we get

$$\begin{aligned}\|x^{k+1} - x^*\| = \|x^k + d^k - x^*\| \; &\leq \; \|d^k - d^k_{ex}\| + \|x^k + d^k_{ex} - x^*\| \\ &\leq \big(C + \frac{1}{\mu}C'\big)\eta_k\|r(x^k)\| + \frac{1}{\mu}\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^* - x^k)\| \\ &\quad + \frac{1}{\mu}\big\|\big(H_k - \nabla^2 f(x^*)\big)d^k\big\|. \end{aligned} \tag{4.17}$$

The twice continuous differentiability of $f$ yields that the second term is $o(\|x^k - x^*\|)$ for $k \to \infty$. The Dennis-Moré condition implies that the third term is $o(\|d^k\|)$. We use Propositions 4.5 and 4.12(a) to get

$$\|d^k\| \leq \|d^k - d^k_{ex}\| + \|d^k_{ex}\| \leq C\eta_k\|r(x^k)\| + \kappa\|x^k - x^*\| \leq \big(C\eta_k(2+L)+\kappa\big)\|x^k - x^*\|. \tag{4.18}$$

Hence, the third term has order $o(\|x^k - x^*\|)$. Thus, the above and Proposition 4.12(a) yield part (c) for $\overline{\eta} = 1/((C_1 + \frac{1}{\mu}C_2)(L+2))$. Finally, under the assumptions of part (d), the first term is also of size $o(\|x^k - x^*\|)$, which completes the proof. □

We close this section with a result on local quadratic convergence under slightly stronger assumption as in Theorem 4.13 (d), in particular, using a stronger version of the Dennis-Moré condition. This condition holds especially for $H_k = \nabla^2 f(x^k)$ or, more generally, $H_k - \nabla^2 f(x^k) = \mathcal{O}(\|d^k\|)$.

**Theorem 4.14.** *Consider Algorithm 4.1 and assume that the sequence $\{H_k\}$ satisfies $mI \preceq H_k \preceq MI$ for all $k \geq 0$ with suitable $0 < m \leq M$ and*

$$\limsup_{k\to\infty} \frac{\big\|\big(H_k - \nabla^2 f(x^*)\big)d^k\big\|}{\|d^k\|^2} < +\infty.$$

*Let $x^*$ be an accumulation point of a sequence $\{x^k\}$ generated by Algorithm 4.1 such that $\psi$ is locally strongly convex in a neighbourhood of $x^*$. If there exists $\kappa' > 0$ such that $\eta_k \leq \kappa' \|r(x^k)\|$ for all $k \geq 0$, the sequence $\{x^k\}$ converges quadratically to $x^*$.*

*Proof.* A closer look at the estimate (4.17) shows that we need to prove

$$\big(C + \frac{1}{\mu}C'\big)\eta_k\|r(x^k)\| + \frac{1}{\mu}\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^* - x^k)\| \\ + \frac{1}{\mu}\big\|\big(H_k - \nabla^2 f(x^*)\big)d^k\big\| \leq K\|x^k - x^*\|^2$$

for some $K > 0$. Choosing $\varepsilon > 0$ sufficiently small such that $\nabla^2 f$ is Lipschitz continuous in $B_\varepsilon(x^*)$, the second term is $\mathcal{O}(\|x^k - x^*\|^2)$. Furthermore, the modified Dennis-Moré-condition above yields that the third term is of size $\mathcal{O}(\|d^k\|^2)$, and (4.18) shows that this is also $\mathcal{O}(\|x^k - x^*\|^2)$. Thus, the claim follows since $\|r(x^k)\| \leq (2 + L)\|x^k - x^*\|$ using Proposition 4.12(a). □

# Chapter 5

# A Regularized Proximal Quasi-Newton Method

We continue to consider solving the optimization problem (1.1) by using proximal-type methods. Unlike the previous discussion, we assume in this section that the convex, proper, lower semicontinuous function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is real valued, while $f : \mathbb{R}^n \to \mathbb{R}$ is still assumed to be at least continuously differentiable. This requirement on $\varphi$ seems quite restrictive, but the resulting class of functions still covers a huge amount of applications, for which the method deduced in this chapter can be applied.

As already seen in the previous chapters, the crucial point of a proximal-type method is the choice of the matrix $H_k$ in the arising subproblems (1.2). In contrast to the earlier methods, the idea of the regularized proximal quasi-Newton method presented in the following is to regularize an approximation $H_k$ of the (possibly not existing) Hessian $\nabla^2 f(x^k)$ by adding the matrix $\mu_k I$ for some $\mu_k > 0$, which results in subproblems of the form

$$\arg\min_{d} \left\{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T (H_k + \mu_k I) d + \varphi(x^k + d) \right\}. \tag{5.1}$$

Depending on the merit of the iteration, the regularization parameter $\mu_k$ is increased or decreased in each iteration according to a trust-region-type scheme. A consequence of this approach is that no classical line search approach, e.g. of Armijo type, is required, which turns out to be quite efficient in numerical examples.

The idea of combining regularization and (proximal) quasi-Newton techniques traces back to the corresponding methods for smooth problems ($\varphi \equiv 0$), where the subproblem (1.2) reduces to $H_k d = -\nabla f(x^k)$, at least if $H_k$ is positive semidefinite. Some improvements [96,147,157,158] have been made similar to our approach in this case. Moreover, trust-region methods for nonsmooth problems in the form (1.1) are considered in various manuscripts [45,61,86,137]. Techniques for the regularization of proximal quasi-Newton methods are investigated in different variations in literature. The proximal Newton method by Lee, Sun, Saunders [95] does not explicitly use a regularization parameter, but the application to proximal quasi-Newton methods covers this idea if the regularization parameter tends to zero. A similar argument holds for the globalized proximal Newton method in Chapter 4. Regularization of $H_k$ by adding a positive multiple of the identity matrix is used in [68,145], but convergence is only shown for convex functions $f$. Approaches for solving the subproblems inexactly in this context are investigated in [92,167]. We outline the main differences of these methods to the presented one subsequent to the detailed description of our algorithm.

The chapter is organized as follows. After introducing the regularized proximal quasi-Newton method itself in Section 5.1, we discuss some of its properties. Quite nice global convergence result in the trust-region framework under mild assumptions are investigated in Section 5.2, followed by the analysis of the convergence using an error bound condition in Section 5.3. In addition, the analysis of local convergence and the superlinear convergence rate is addressed in Section 5.4. Nevertheless, although covering this issue, a drawback of the method is that the subproblem (5.1) is sometimes not solvable. To address this, we also present a modified version of the algorithm in Section 5.5 which combines the method with a proximal gradient framework similar to the approach in Chapter 4. Although theoretically not providing any theoretical advantages, this modification can significantly improve the speed of convergence in numerical examples.

We note that an essential part of the Sections 5.1-5.3 mainly coincide with the preprint [81].

## 5.1 Algorithmic Framework

This section contains a detailed derivation and discussion of the regularized proximal quasi-Newton method. The method combines two main concepts: First, solving the regularized problem (5.1) instead of (1.2) to compute a suitable search direction $d^k$, and second, the appropriate choice of the regularization parameter $\mu_k$ according to a trust-region-type framework. An advantage of the regularization is that subproblems are more likely to have a solution, if the matrix $H_k$ is not positive semidefinite.

Provided that the computed search direction $d^k$ is accepted by an appropriate backtracking strategy in order to guarantee global convergence, the next iterate in general proximal-type methods is set to $x^{k+1} := x^k + d^k$. Here, the globalization is achieved by the adjustment of the regularization parameter $\mu_k > 0$, no classical line search is required (which might result in many function evaluations), and no trust-region radius is needed (in particular, no trust-region-type subproblem has to be solved). Instead, however, additional evaluations of the proximity operator might be necessary, which can be quite expensive computationally. Nevertheless, numerical tests show that this additional effort leads to significantly less iterations and thus lower overall costs, and on the other hand, trust-region-type methods are very appropriate, especially for nonconvex global optimization problems.

The regularized proximal quasi-Newton method therefore considers the regularized approximation

$$\hat{q}_k(d) := f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T (H_k + \mu_k I) d + \varphi(x^k + d) \tag{5.2}$$

and the search direction $d^k$ is computed as a minimizer thereof. Furthermore, let $q_k(d) := f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d + \varphi(x^k + d)$ be the corresponding approximation without regularization by $\mu_k$. To control the success of a candidate $d^k$ we define the *predicted reduction* of $\psi$ in step $k \geq 0$ via

$$\mathrm{pred}_k := q_k(0) - q_k(d^k) = -\big(\nabla f(x^k)^T d^k + \varphi(x^k + d^k) - \varphi(x^k)\big) - \tfrac{1}{2}(d^k)^T H_k d^k$$

and the *actual reduction* of $\psi$ as

$$\mathrm{ared}_k := \psi(x^k) - \psi(x^k + d^k).$$

Note that $\mathrm{pred}_k = -\Delta_k - \frac{1}{2}(d^k)^T H_k d^k$, where $\Delta_k$ is defined as in Algorithms 3.1 and 4.1. The ratio $\rho_k := \mathrm{ared}_k \,/\, \mathrm{pred}_k$ of these quantities is, similar to trust-region methods [51], used to control the update of the regularization parameter and the iterate. Since $H_k$ does not need to be positive definite, we must take into account that a minimizer of $\hat{q}_k$ may not exist or the corresponding value $\mathrm{pred}_k$ is not (sufficiently) positive. These situations are handled as unsuccessful steps and the quality of the update is controlled by the sufficient decrease criterion stated in (5.3). For this purpose, similar to Chapter 4, we set

$$r_H(x) := \operatorname{prox}_{\varphi}^{H}\left(x - H^{-1}\nabla f(x)\right) - x = \operatorname*{arg\,min}_{d\in\mathbb{R}^n} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + \varphi(x+d) \right\}.$$

for $x \in \mathbb{R}^n$ and $H \in \mathbb{S}^n_{++}$, as well as $r(x) := r_I(x)$. Altogether, this motivates Algorithm 5.1.

**Algorithm 5.1** Regularized Proximal Quasi-Newton Method (RPQN)

```
(S.0) Choose $x^0 \in \mathbb{R}^n$, parameters $\mu_0 > 0$, $p_{\min} \in (0, 1/2)$, $\kappa > 1$, $c_1 \in (0, 1/2)$,
      $c_2 \in (c_1, 1)$, $\sigma_1 \in (0, 1), \sigma_2 > 1$, and set $k := 0$.
(S.1) If $x^k$ satisfies a suitable termination criterion: STOP.
(S.2) Choose $H_k \in \mathbb{S}^n$, and find a solution $d^k$ of the problem
          $\min_d \hat{q}_k(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T (H_k + \mu_k I) d + \varphi(x^k + d).$
      If this problem has no solution, or if
          $\mathrm{pred}_k \le p_{\min} \|d^k\| \cdot \min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\},$      (5.3)
      set $x^{k+1} := x^k$, $\mu_{k+1} := \sigma_2 \mu_k$, and go to (S.4). Otherwise go to (S.3).
(S.3) Set $\rho_k := \mathrm{ared}_k \,/\, \mathrm{pred}_k$ and perform the following updates:
          $x^{k+1} := \begin{cases} x^k & \text{if } \rho_k \le c_1, \\ x^k + d^k & \text{otherwise,} \end{cases}$
          $\mu_{k+1} := \begin{cases} \sigma_2 \mu_k & \text{if } \rho_k \le c_1, \\ \mu_k & \text{if } c_1 < \rho_k \le c_2, \\ \sigma_1 \mu_k & \text{otherwise.} \end{cases}$
(S.4) Update $k \leftarrow k + 1$, and go to (S.1).
```

In the following discussion, we call an iteration $k$

- *unsuccessful*, if (S.3) is skipped or $\rho_k \le c_1$,
- *successful,* if $c_1 < \rho_k \le c_2$,
- *highly successful,* if $\rho_k > c_2$.

Note that, in an unsuccessful iteration, both (S.2) and (S.3) keep the current iterate $x^k$ and choose a larger regularization parameter. In all other iterations, we perform the update $x^{k+1} = x^k + d^k$ and either keep the regularization parameter $\mu_k$ (in successful iterations) or reduce this parameter (in highly successful iterations). We also stress that a test like (5.3) is not required by trust-region methods since, there, the corresponding predicted reduction is automatically positive, whereas this cannnot be guaranteed in our setting. Whenever we reach (S.3), however, the value of $\mathrm{pred}_k$ is (sufficiently) positive, which, in turn, implies that the overall method is well-defined.

Furthermore, we note that in contrast to the corresponding method in [81], the test (5.3) was slightly adapted to lay the foundations for the local convergence analysis. In addition, we assume $c_1 < 1/2$ instead of $c_1 < 1$ here, which would be the natural limit for trust-region-type methods. This is necessary for the results in Section 5.3 and not at all a restriction in practice, since the aim is to choose $c_1$ small.

We briefly discuss the differences between Algorithm 5.1 and some affiliated methods. The methods in [68,145] are based on a similar regularization than ours, where the regularization parameter is only increased if a suitable criterion is not satisfied for the solution of the subproblems. In contrast to our method, they do not consider the possibility to reduce the regularization parameter if an iterate is highly successful. Convergence is shown under the assumption of strong convexity of $f$. Furthermore, they combine the method with an inexactness criterion for the subproblem and use a FISTA-type acceleration. In this case, a main assumption on $f$ is convexity.

The inexact algorithms by Lee and Wright [92] use two different types of regularization: $H_k + \mu_k I$ or $\mu_k \cdot H_k$ with a positive regularization parameter $\mu_k$, which is initially set to 1 in each step and increased until a sufficient decrease condition is satisfied. In contrast to our method, it is not possible to choose $\mu_k$ small if the iterate is close to a solution. Convergence is shown for $\nabla f$ being Lipschitz continuous (but $f$ not necessarily convex). Moreover, some improved convergence results are provided for strongly convex functions.

Yue et al. [167] developed another inexact regularized proximal Newton method. A main difference to our approach is that, instead of an approximation $H_k$, the exact Hessian of $f$ is used and the regularization parameter $\mu_k$ is chosen due to the optimality of the current iterate, and not based on the quality of the current update. Furthermore, the subproblems are solved inexactly, and an Armijo-type line search is needed. The convergence proof requires convexity of $f$ and uses an error bound.

In contrast to these methods, we do not provide a theory for inexact solutions of the subproblems in (S.2). It turns out that this is not necessary since these problems can be solved very efficiently and with high accuracy in our numerical examples using Algorithm 3.2.

**Remark 5.1** (Termination criterion)**.** In view of Proposition 3.7, we know that $x^k$ is a stationary point of $\psi$ if and only if $r(x^k) = 0$. Combining this property with the (uniform) continuity of $r(\cdot)$ yields an appropriate termination criterion for Algorithm 5.1. Since $x^k$ is a stationary point if $d^k = 0$, cf. again Proposition 3.7, the method is well-defined. However, the norm of the search direction $d^k$ is not a good choice for a termination criterion, since $x^k$ might be a stationary point without having $d^k = 0$, if $H_k + \mu_k I$ is not positive definite. $\Diamond$

## 5.2 Global Convergence in a Trust-Region Framework

In this section, we investigate the global convergence properties of Algorithm 5.1. To this end, we assume that Algorithm 5.1 generates an infinite sequence $\{x^k\}$ and does not terminate after finitely many steps. Though, formally, we did not specify the termination criterion in (S.1), any suitable stopping criterion will include a test whether the current point $x^k$ is already a stationary point of the given optimization problem, cf. Remark 5.1. To simplify some of the subsequent phrases, we therefore assume throughout this section that none of the iterates $x^k$ is already a stationary point. Then, by Proposition 3.7, we have $d^k \neq 0$ for all $k$.

Note that the subsequent global convergence analysis of Algorithm 5.1 does not require the matrices $H_k$ to be good approximations of the corresponding (possibly not existing) Hessians $\nabla^2 f(x^k)$. We only need that the sequence $\{H_k\}$ is bounded. Before presenting the two main global convergence theorems, we establish some technical results.

**Lemma 5.2.** *Let $\{H_k\}$ be a bounded sequence of symmetric matrices. Assume that $\mu_k \to \infty$ and $\{x^k\} \subset \mathbb{R}^n$ converges to a nonstationary point of $\psi$. Then*

$$\limsup_{k\to\infty} \frac{\|r(x^k)\|}{\|r_{H_k+\mu_k I}(x^k)\| \cdot \mu_k} \leq 1.$$

*Proof.* The assumptions imply that $H_k + \mu_k I$ is positive definite for all sufficiently large $k$ and $r_{H_k+\mu_k I}(x^k)$ is therefore well-defined. Furthermore, $\|r_{H_k+\mu_k I}(x^k)\| \neq 0$ for sufficiently large $k$ since the limit point $\overline{x}$ of $\{x^k\}$ is not a stationary point of $\psi$. Thus, we apply Lemma 3.6 with $H = H_k + \mu_k I$ and $\widehat{H} = I$ to get

$$\frac{\|r(x^k)\|}{\|r_{H_k+\mu_k I}(x^k)\|} \leq \Big(1 + \frac{1}{\lambda_{\min}(H_k) + \mu_k}\Big) \cdot \big(\lambda_{\max}(H_k) + \mu_k\big).$$

Dividing this estimate by $\mu_k$, using the boundedness of the sequence $\{H_k\}$, and taking the limit $k \to \infty$, it follows that the expression on the right-hand side tends to one, which yields the claim. □

Recall that if $H_k + \mu_k I$ is positive definite, the step $d^k$ can equivalently be written as $d^k = r_{H_k+\mu_k I}(x^k)$. In the next result, we show that the corresponding sequence $\{d^k\}$ is a vanishing sequence under the assumptions that the sequence $\{\mu_k\}$ tends to $+\infty$ and $\{x^k\}$ is a bounded sequence.

**Proposition 5.3.** *Let $\{H_k\}$ be a bounded sequence of symmetric matrices. Assume that $\mu_k \to \infty$ and the sequence $\{x^k\} \subset \mathbb{R}^n$ generated by Algorithm 5.1 is bounded. Let $d^k := r_{B_k+\mu_k I}(x^k)$. Then $d^k \to 0$.*

*Proof.* Note that the boundedness of the sequence $\{H_k\}$ and $\mu_k \to \infty$ imply that $d^k$ is well-defined for sufficiently large $k$. Moreover, the definition of successful and highly successful steps connotes that the sequence $\{\psi(x^k)\}$ is a nonincreasing sequence. Let $M > 0$ and $k_0 \in \mathbb{N}$ such that $H_k + \mu_k I \succeq MI$ holds for all $k \geq k_0$. Furthermore, let $\mathcal{X} \subset \mathbb{R}^n$ be a compact set such that $x^k \in \mathcal{X}$ holds for all $k \geq k_0$, and $\bar{u}^k \in \partial\varphi(x^k)$. Then, for any $k \geq k_0$ we get

$$\begin{aligned}
\psi(x^0) &\geq \psi(x^k) = \hat{q}_k(0) \\
&\geq \hat{q}_k(d^k) = f(x^k) + \nabla f(x^k)^T d^k + \frac{1}{2}(d^k)^T(H_k + \mu_k I)d^k + \varphi(x^k + d^k) \\
&\geq f(x^k) + \nabla f(x^k)^T d^k + \frac{M}{2}\|d^k\|^2 + \varphi(x^k) + (\bar{u}^k)^T d^k \\
&\geq \min_{x \in X} \psi(x) - \|d^k\| \max_{x\in\mathcal{X}, u\in\partial\varphi(x)} \|\nabla f(x) + u\| + \frac{M}{2}\|d^k\|^2 \\
&= \alpha_1 - \alpha_2\|d^k\| + \frac{M}{2}\|d^k\|^2 =: \tilde{q}(d^k),
\end{aligned}$$

where $\alpha_1 := \min_{x\in\mathcal{X}} \psi(x)$ and $\alpha_2 := \max_{x\in\mathcal{X},u\in\partial\varphi(x)} \|\nabla f(x) + u\|$. Hence, $\tilde{q}$ is coercive. In particular, the level set $\operatorname{lev}_{\leq\psi(x^0)} \tilde{q}$ is bounded (Proposition 2.15), and therefore the sequence $\{d^k\}$ is bounded.
By definition of $d^k$ and Fermat's rule (Proposition 2.47) applied to $\hat{q}_k$, there exists $u^k \in \partial\varphi(x^k + d^k)$ such that

$$0 = \nabla f(x^k) + u^k + H_k d^k + \mu_k d^k.$$

As the sequences $\{\nabla f(x^k)\}$, $\{u^k\}$ (cf. Proposition 2.24) and $\{H_k d^k\}$ are bounded, the sequence $\{\mu_k d^k\}$ must also be bounded. From $\mu_k \to \infty$, we get $d^k \to 0$. □

The following result will be applied to the situation where we have only finitely many successful iterations, i.e. where $x^k$ stays constant eventually, say $x^k = \overline{x} \in \mathbb{R}^n$ for all $k$ sufficiently large. By the assumption that Algorithm 5.1 generates an infinite sequence, this means that $\overline{x}$ is a nonstationary point of $\psi$. To avoid any ambiguity in the notation, we write $\bar{d}^k := r_{H_k+\mu_k I}(\overline{x})$, although, in the subsequent application, we will eventually have $\bar{d}^k = d^k$ since $\overline{x}$ corresponds to $x^k$ for all sufficiently large $k$.

**Lemma 5.4.** *Let $\{H_k\}$ be a bounded sequence of symmetric matrices. Assume that $\mu_k \to \infty$ and $\overline{x}$ is a nonstationary point of $\psi$. Define $\bar{d}^k := r_{H_k+\mu_k I}(\overline{x})$, and let $s$ be an accumulation point of the sequence $\{\bar{d}^k / \|\bar{d}^k\|\}$. Then $\psi'(\overline{x}; s) < 0$.*

*Proof.* Using the previous result, we get $\bar{d}^k \to 0$. Furthermore, applying Fermat's rule to $\hat{q}_k$, we obtain

$$0 = \nabla f(\overline{x}) + (H_k + \mu_k I)\bar{d}^k + u^k \tag{5.4}$$

for some $u^k \in \partial\varphi(\overline{x} + \bar{d}^k)$. The boundedness of the subdifferential (Proposition 2.24) yields that the sequence $\{u^k\}$ is bounded. Thus, we can choose a subsequence $\mathcal{K} \subset \mathbb{N}$ such that

$$\frac{\bar{d}^k}{\|\bar{d}^k\|} \to_{\mathcal{K}} s \qquad \text{and} \qquad u^k \to_{\mathcal{K}} \overline{u}.$$

The closedness of the subdifferential (Proposition 2.23) yields $\overline{u} \in \partial\varphi(\overline{x})$. By assumption, we therefore have $\nabla f(\overline{x}) + \overline{u} \neq 0$. Furthermore, using Lemma 3.9, we obtain

$$\psi'(\overline{x}, \bar{d}^k) \leq -(\bar{d}^k)^T (H_k + \mu_k I)\bar{d}^k \leq -\big(\lambda_{\min}(H_k) + \mu_k\big)\|\bar{d}^k\|^2.$$

Since (5.4) implies $\|\nabla f(\overline{x}) + u^k\| = \|(H_k + \mu_k I)\bar{d}^k\| \leq (\|H_k\| + \mu_k)\|\bar{d}^k\|$, we get

$$\psi'(\overline{x}, \bar{d}^k) \leq -\big(\lambda_{\min}(H_k) + \mu_k\big)\|\bar{d}^k\|^2 \leq -\|\nabla f(\overline{x}) + u^k\| \cdot \frac{\lambda_{\min}(H_k) + \mu_k}{\|H_k\| + \mu_k} \cdot \|\bar{d}^k\|.$$

Thus, the sublinearity of $\psi'(\overline{x}, \cdot)$ yields

$$\psi'\left(\overline{x}, \frac{\bar{d}^k}{\|\bar{d}^k\|}\right) \leq -\|\nabla f(\overline{x}) + u^k\| \cdot \frac{\lambda_{\min}(H_k) + \mu_k}{\|H_k\| + \mu_k}.$$

For $k \in \mathcal{K}, k \to \infty$, the right-hand side converges to $-\|\nabla f(\overline{x}) + \overline{u}\|$. Since $\varphi$ is real-valued, the directional derivative $\psi'(\overline{x}, \cdot)$ is continuous, and we obtain

$$\psi'(\overline{x}, s) = \lim_{k \in \mathcal{K}, k \to \infty} \psi'\Big(\overline{x}, \frac{\bar{d}^k}{\|\bar{d}^k\|}\Big) \leq -\|\nabla f(\overline{x}) + \overline{u}\| < 0,$$

which completes the proof. □

We now apply the previous result to show that there always exist infinitely many successful or highly successful iterations.

**Lemma 5.5.** *Let $\{H_k\}$ be a bounded sequence of symmetric matrices. Then Algorithm 5.1 performs infinitely many successful or highly successful steps.*

*Proof.* We follow the corresponding proof of [147] and assume, by contradiction, that there exists $k_0 \in \mathbb{N}$ such that all steps $k \geq k_0$ are unsuccessful. This implies $x^k = x^{k_0}$ for all $k \geq k_0$ and, due to the implicit assumption that Algorithm 5.1 generates an infinite sequence, that $\mu_k \to +\infty$. Since $\{H_k\}$ is a bounded sequence, the matrices $H_k + \mu_k I$ are therefore positive definite for all sufficiently large $k \geq k_0$. In view of Proposition 3.7 and $d^k \neq 0$ (otherwise we would have stopped after finitely many iterations), it follows that $x^{k_0}$ is a nonstationary point of $\psi$ and $r(x^{k_0}) \neq 0$. Moreover, the positive definiteness of

$H_k + \mu_k I$ guarantees that the search directions $d^k$ are well-defined for $k \geq k_0$. Recalling that $p_{\min} < \frac{1}{2}$ and $d^k = r_{H_k+\mu_k I}(x^k)$, we have

$$\frac{\|r(x^k)\|}{\|d^k\|\mu_k} < \frac{1}{2p_{\min}}$$

for all sufficiently large $k \geq 0$ in view of Lemma 5.2. Then, by using $\hat{q}_k(d^k) \leq \hat{q}_k(0)$, we obtain

$$\begin{aligned}
\text{pred}_k &= \psi(x^k) - q_k(d^k) \\
&= \psi(x^k) - \hat{q}_k(d^k) + \frac{\mu_k}{2}\|d^k\|^2 \\
&\geq \psi(x^k) - \hat{q}_k(0) + \frac{\mu_k}{2}\|d^k\|^2 = \frac{\mu_k}{2}\|d^k\|^2 \\
&> p_{\min}\|r(x^k)\| \cdot \|d^k\| \geq p_{\min}\|d^k\| \cdot \min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\}. \qquad (5.5)
\end{aligned}$$

Hence, for all sufficiently large $k \geq 0$, Algorithm 5.1 performs (S.3). Since all iterations $k \geq k_0$ are unsuccessful, this means $\text{ared}_k \leq c_1 \, \text{pred}_k$. It follows that

$$\psi(x^{k_0} + d^k) - \psi(x^{k_0}) \geq c_1\big(\nabla f(x^{k_0})^T d^k + \varphi(x^{k_0} + d^k) - \varphi(x^{k_0}) + \tfrac{1}{2}(d^k)^T H_k d^k\big).$$

for $k \geq k_0$, possibly after enlarging $k_0$. Setting $t_k = \|d^k\|$ and dividing this estimate by $t_k$ yields

$$\begin{aligned}
&\frac{\psi\big(x^{k_0} + t_k\frac{d^k}{\|d^k\|}\big) - \psi(x^{k_0})}{t_k} \\
&\qquad \geq c_1\left(\nabla f(x^{k_0})^T\frac{d^k}{\|d^k\|} + \frac{\varphi\big(x^{k_0} + t_k\frac{d^k}{\|d^k\|}\big) - \varphi(x^{k_0})}{t_k} + \frac{1}{2}\frac{(d^k)^T}{\|d^k\|}H_k d^k\right).
\end{aligned}$$

Using the convergence $\{d^k\} \to 0$, see Proposition 5.3, the boundedness of $\{H_k\}$, and choosing a subsequence $\mathcal{K}$ such that $d^k/\|d^k\| \to s$, taking the limit on this sequence yields $\psi'(x^{k_0}; s) \geq c_1\psi'(x^{k_0}; s)$, where we used Proposition 2.38. Since $c_1 \in (0, 1/2)$, this results in $\psi'(x^{k_0}; s) \geq 0$, a contradiction to Lemma 5.4. Thus, there are infinitely many successful or highly successful iterations. □

We next formulate two global convergence results. The corresponding statements are similar to those known for trust-region methods in unconstrained optimization.

**Theorem 5.6.** *Let $\{H_k\}$ be a bounded sequence of symmetric matrices, and assume that $\psi$ is bounded from below. Then any sequence $\{x^k\}$ generated by Algorithm 5.1 satisfies* $\liminf_{k\to\infty} \|r(x^k)\| = 0$.

*Proof.* Let $\mathcal{S} \subset \mathbb{N}$ be the (infinite) set of successful and highly successful iterations. Contrary to the claim, assume that $\liminf_{k\to\infty} \|r(x^k)\| > 0$. Then there exists $k_0 \in \mathbb{N}$ and $\varepsilon > 0$ such that $\min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\} \geq \varepsilon$ for all $k \geq k_0$. By the definition of successful steps, we get

$$\psi(x^k) - \psi(x^{k+1}) \geq c_1 \, \text{pred}_k \geq p_{\min}c_1\|d^k\| \cdot \min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\} \geq p_{\min}c_1\varepsilon\|d^k\|$$

for all $k \in \mathcal{S}, k \geq k_0$. Since $\psi$ is bounded from below, summation yields

$$\infty > \sum_{k=0}^{\infty} \big[\psi(x^k) - \psi(x^{k+1})\big] = \sum_{k \in \mathcal{S}} \big[\psi(x^k) - \psi(x^{k+1})\big] \geq p_{\min} c_1 \varepsilon \sum_{k \in \mathcal{S}} \|d^k\|.$$

Taking into account that $x^k$ is not updated in unsuccessful steps, it follows that

$$\infty > \sum_{k \in \mathcal{S}} \|d^k\| = \sum_{k \in \mathcal{S}} \|x^{k+1} - x^k\| = \sum_{k=0}^{\infty} \|x^{k+1} - x^k\|.$$

Hence, $\{x^k\}$ is a Cauchy sequence and therefore convergent to some $\overline{x} \in \mathbb{R}^n$. Since $\|r(\overline{x})\| = \lim_{k \to \infty} \|r(x^k)\| \geq \varepsilon$, $\overline{x}$ is not a stationary point of $\psi$.
By Lemma 5.5, there are infinitely many successful or highly successful steps and, as shown above, we have $\|d^k\| \to_{\mathcal{S}} 0$. Similar to (5.4), there holds

$$0 = \nabla f(x^k) + (H_k + \mu_k I)d^k + u^k$$

for some $u^k \in \partial\varphi(x^k + d^k)$. Assuming that the sequence $\{\mu_k\}_{\mathcal{S}}$ is bounded, $(H_k + \mu_k I)d^k$ converges to 0 for $k \in \mathcal{S}, k \to \infty$. Furthermore, Proposition 2.24 yields that $\{u^k\}_{\mathcal{S}}$ is bounded and we can choose a subsequence $\mathcal{K} \subset \mathcal{S}$ such that $u^k \to_{\mathcal{K}} \overline{u}$ with $\overline{u} \in \partial\varphi(\overline{x})$. As a consequence, taking the limit $k \in \mathcal{K}, k \to \infty$ in the above equation yields $0 = \nabla f(\overline{x}) + \overline{u} \in \nabla f(\overline{x}) + \partial\varphi(\overline{x})$, in contradiction to the nonstationarity of $\overline{x}$.
Hence, without loss of generality, we have $\{\mu_k\}_{\mathcal{S}} \to \infty$. It follows that $\{\mu_k\} \to \infty$ since $\mu_k$ cannot decrease during unsuccessful iterations. This implies that Algorithm 5.1 also performs infinitely many unsuccessful iterations. On the other hand, in the same way as in (5.5), we get

$$\psi(x^k) - q_k(d^k) = \mathrm{pred}_k \geq p_{\min}\|d^k\| \cdot \min\{\|r(x^k)\|, \|r(x^k)\|^{\kappa}\} \geq p_{\min}\varepsilon\|d^k\|$$

for sufficiently large $k \geq 0$. For every such $k$, there exists $\xi^k$ on the straight line between $x^k$ and $x^k + d^k$ such that $f(x^k + d^k) - f(x^k) = \nabla f(\xi^k)^T d^k$. By the convergence of $\{x^k\}$ to $\overline{x}$ and since $\{d^k\} \to 0$ in view of Proposition 5.3, the sequence $\{\xi^k\}$ also converges to $\overline{x}$. Thus, we obtain

$$\begin{aligned}
\big|\rho_k - 1\big| &= \Big|\frac{\mathrm{ared}_k}{\mathrm{pred}_k} - 1\Big| = \Big|\frac{\psi(x^k) - \psi(x^k + d^k)}{\psi(x^k) - q_k(d^k)} - 1\Big| \\
&= \Big|\frac{\psi(x^k + d^k) - q_k(d^k)}{\psi(x^k) - q_k(d^k)}\Big| \\
&\leq \frac{1}{p_{\min}\varepsilon} \, \frac{\big|f(x^k + d^k) - f(x^k) - \nabla f(x^k)^T d^k\big| + \frac{1}{2}\big|(d^k)^T H_k d^k\big|}{\|d^k\|} \\
&= \frac{1}{p_{\min}\varepsilon} \, \frac{\big|\nabla f(\xi^k)^T d^k - \nabla f(x^k)^T d^k\big|}{\|d^k\|} + \frac{1}{2p_{\min}\varepsilon}\Big|(d^k)^T H_k \frac{d^k}{\|d^k\|}\Big| \\
&\leq \frac{1}{p_{\min}\varepsilon} \, \big\|\nabla f(\xi^k) - \nabla f(x^k)\big\| + \frac{1}{p_{\min}\varepsilon} \, \|H_k\| \cdot \|d^k\| \longrightarrow 0
\end{aligned}$$

for $k \to \infty$. Hence, $\{\rho_k\} \to 1$, i.e. eventually all steps are highly successful, which yields a contradiction. □

Similar to trust-region methods, the previous result can be used to prove a stronger statement for functions with uniformly continuous gradient. The proof generalizes the one of [147, Theorem 3.5].

**Theorem 5.7.** *Let $\{H_k\}$ be a bounded sequence of symmetric matrices, assume that $\psi$ is bounded from below and that $\nabla f$ is uniformly continuous on a set $\mathcal{X}$ satisfying $\{x^k\} \subset \mathcal{X}$, where $\{x^k\}$ denotes a sequence generated by Algorithm 5.1. Then*

$$\lim_{k \to \infty} \|r(x^k)\| = 0;$$

*in particular, every accumulation point of the sequence $\{x^k\}$ is a stationary point of $\psi$.*

*Proof.* Assume, by contradiction, that there exists $\delta > 0$ and $\mathcal{K} \subset \mathbb{N}$ such that $\|r(x^k)\| \geq 2\delta$ holds for all $k \in \mathcal{K}$. Set $\overline{\delta} := \min\{\delta, \delta^\kappa\}$. By Theorem 5.6, for each $k \in \mathcal{K}$, there is an index $\ell(k) > k$ such that $\|r(x^l)\| \geq \delta$ holds for all $k \leq l < \ell(k)$ and $\|r(x^{\ell(k)})\| < \delta$.
If, for $k \in \mathcal{K}$, an iteration $k \leq l < \ell(k)$ is successful or highly successful, we get

$$\psi(x^l) - \psi(x^{l+1}) \geq c_1 \operatorname{pred}_l \geq c_1 p_{\min} \min\{\|r(x^l)\|, \|r(x^l)\|^\kappa\} \cdot \|d^l\| \geq c_1 p_{\min} \overline{\delta} \|x^{l+1} - x^l\|.$$

For unsuccessful iterations $l$, this estimate holds trivially. Thus,

$$p_{\min} c_1 \overline{\delta} \|x^{\ell(k)} - x^k\| \leq p_{\min} c_1 \overline{\delta} \sum_{l=k}^{\ell(k)-1} \|x^{l+1} - x^l\| \leq \sum_{l=k}^{\ell(k)-1} \psi(x^l) - \psi(x^{l+1}) = \psi(x^k) - \psi(x^{\ell(k)})$$

holds for all $k \in \mathcal{K}$. By assumption, $\psi$ is bounded from below, and by construction, the sequence $\{\psi(x^k)\}$ is monotonically decreasing, hence convergent. This implies that the sequence $\big\{\psi(x^k) - \psi(x^{\ell(k)})\big\}_{\mathcal{K}}$ converges to zero. Hence, we get $\big\{\|x^{\ell(k)} - x^k\|\big\} \to_{\mathcal{K}} 0$. The uniform continuity of $\nabla f$ and of the proximity operator (Proposition 3.5) together with the fact that the composition of uniformly continuous functions is uniformly continuous, yields the uniform continuity of the residual function $r(\cdot)$. Thus, we get $\big\{\|r(x^{\ell(k)}) - r(x^k)\|\big\} \to_{\mathcal{K}} 0$. On the other hand, by the choice of $\ell(k)$, we have

$$\big\|r(x^k) - r(x^{\ell(k)})\big\| \geq \big\|r(x^k)\big\| - \big\|r(x^{\ell(k)})\big\| \geq 2\delta - \delta \geq \delta,$$

which yields the desired contradiction. □

## 5.3 Convergence Using an Error Bound Condition

In the following we deduce further convergence results for the regularized proximal quasi-Newton method in Algorithm 5.1. To this end, we first provide some useful and path-breaking results and, if $\nabla f$ is Lipschitz continuous, prove the boundedness of the sequence $\{\mu_k\}$ of regularizers. Moreover, we introduce an error bound condition to end up with the convergence of the entire sequence. As in the previous section, we assume that Algorithm 5.1 generates an infinite sequence of iterates $\{x^k\}$ and does not terminate with a stationary point of $\psi$. We start with some technical estimates.

**Lemma 5.8.** *Assume that the sequence $\{H_k\}$ is uniformly bounded and positive definite, i.e. there exist constants $0 < m \leq M$ such that $mI \preceq H_k \preceq MI$ holds for all $k \geq 0$. Then the following estimates hold with the notation of Algorithm 5.1:*

*(a)* $\operatorname{pred}_k \geq \dfrac{1}{2}(m + 2\mu_k)\|d^k\|^2$,

*(b)* $\dfrac{\|r(x^k)\|}{\|d^k\|} \leq \Big(1 + \dfrac{1}{m + \mu_k}\Big)(M + \mu_k) \leq \dfrac{m+1}{m}(M + \mu_k)$,

*(c)* $\dfrac{\|d^k\|}{\|r(x^k)\|} \leq \dfrac{1+M+\mu_k}{m+\mu_k} \leq \dfrac{1+M}{m}$.

*Proof.* (a) Using the second estimate in Lemma 3.9, we get

$$\begin{aligned}\text{pred}_k &= -\big(\nabla f(x^k)^T d^k + \varphi(x^k+d^k) - \varphi(x^k)\big) - \frac{1}{2}(d^k)^T H_k d^k \\ &\geq (d^k)^T(H_k+\mu_k I)d^k - \frac{1}{2}(d^k)^T H_k d^k \geq \frac{1}{2}(m+2\mu_k)\|d^k\|^2.\end{aligned}$$

The estimates in parts (b) and (c) follow directly from Lemma 3.6 using $\lambda_{\max}(H_k+\mu_k I) \leq M+\mu_k$ and $\lambda_{\min}(H_k+\mu_k I) \geq m+\mu_k$. □

The next result is essential to prove the boundedness of the sequence of regularizers $\{\mu_k\}$.

**Lemma 5.9.** *Assume that $\nabla f$ is Lipschitz continuous with Lipschitz constant $L > 0$ and $H_k \succeq mI$ for some $m > 0$. If, for some iteration $k$ in Algorithm 5.1, we have $\mu_k \geq \overline{\mu} := \max\{L-m, 0\}$, there holds*

$$\text{ared}_k > c_1 \,\text{pred}_k\,.$$

*Proof.* Let $\mu_k \geq \overline{\mu}$. Then $H_k+\mu_k I \succeq LI$ and the Lipschitz continuity of $\nabla f$ yields

$$f(x^k+d^k) - f(x^k) \leq \nabla f(x^k)^T d^k + \frac{1}{2}L\|d^k\|^2 \leq \nabla f(x^k)^T d^k + \frac{1}{2}(d^k)^T(H_k+\mu_k I)d^k,$$

which is equivalent to

$$\psi(x^k+d^k) - \psi(x^k) \leq \nabla f(x^k)^T d^k + \varphi(x^k+d^k) - \varphi(x^k) + \frac{1}{2}(d^k)^T(H_k+\mu_k I)d^k.$$

Hence, using the definition of $\text{pred}_k$ and $\text{ared}_k$, we get $-\text{ared}_k \leq -\text{pred}_k + \frac{\mu_k}{2}\|d^k\|^2$. A combination with Lemma 5.8(a) yields

$$\text{ared}_k \geq \text{pred}_k - \frac{\mu_k}{2}\|d^k\|^2 \geq \text{pred}_k \cdot \frac{\mu_k+m}{2\mu_k+m} > \frac{1}{2}\text{pred}_k \geq c_1\,\text{pred}_k,$$

which had to be shown (note that we need $c_1 \leq \frac{1}{2}$ at this point). □

For the boundedness of the sequence $\{\mu_k\}$, it remains to prove that condition (5.3) holds for sufficiently large $\mu_k > 0$, which is the aim of the next result.

**Proposition 5.10.** *Assume that $\nabla f$ is Lipschitz continuous with Lipschitz constant $L > 0$ and $MI \succeq H_k \succeq mI$ for some $M \geq m > 0$. Then, the sequence $\{\mu_k\}$ generated by Algorithm 5.1 is bounded.*

*Proof.* Assume that the sequence $\{\mu_k\}$ is unbounded. This means, there is a subsequence $\mathcal{K} \subset \mathbb{N}_0$ such that $\{\mu_k\}_{\mathcal{K}} \to \infty$. Since $\mu_k$ cannot increase in successful or highly successful steps, this means that there are infinitely many unsuccessful iterations. Without loss of generality we assume that all steps $k \in \mathcal{K}$ are unsuccessful. In view of Lemma 5.9 this is only possible if for all sufficiently large $k \in \mathcal{K}$ we have

$$\text{pred}_k \leq p_{\min}\|d^k\| \cdot \|r(x^k)\|^\kappa,$$

Using Lemma 5.8(a), this yields

$$\frac{m+2\mu_k}{2}\|d^k\| \leq p_{\min}\|r(x^k)\|^\kappa \qquad \Longleftrightarrow \qquad \frac{\|r(x^k)\|}{\mu_k\|d^k\|} \geq \frac{m+2\mu_k}{2p_{\min}\mu_k}\|r(x^k)\|^{1-\kappa}.$$

We combine this estimate with Lemma 5.8(b) to get

$$\left(1+\frac{1}{m+\mu_k}\right)\frac{M+\mu_k}{\mu_k} \geq \frac{m+2\mu_k}{2p_{\min}\mu_k}\|r(x^k)\|^{1-\kappa} \geq \frac{m+2\mu_k}{2p_{\min}\mu_k} \tag{5.6}$$

for sufficiently large $k \in \mathcal{K}$, such that $\|r(x^k)\| \leq 1$, cf. Theorem 5.7. Taking the limit in $\mathcal{K}$, the left hand side of this estimate converges to 1, whereas the right hand side converges to $1/p_{\min} > 1$, which yields a contradiction. Hence, the sequence $\{\mu_k\}$ is bounded. □

We note that this proof does not require $\kappa > 1$. Instead, the complete proof works out assuming $\kappa = 1$. However, as the proof of the local convergence theorem will need (5.6) for $\kappa > 1$, it seemed more reasonable to deduce that estimate already at this point.

To prove the convergence of the complete sequence, we need an additional preliminary. For a long time, the natural assumption to prove local convergence properties and to state convergence rates was strong convexity of the objective function. In recent years this property has frequently been replaced in favor of more general conditions. Here, we assume that $\psi$ satisfies a local error bound condition, which is used by Tseng and Yun in [156].

**Assumption 5.11.** Assume that $\psi$ is bounded from below and $\mathcal{X}^* \neq \emptyset$, where $\mathcal{X}^*$ is the set of stationary points of $\psi$.

(a) For any $\zeta \geq \min_x \psi(x)$, there exist scalars $\tau > 0$ and $\varepsilon > 0$ such that

$$\operatorname{dist}(x, \mathcal{X}^*) \leq \tau\|r(x)\| \quad \text{whenever} \quad \psi(x) \leq \zeta,\ \|r(x)\| \leq \varepsilon.$$

(b) There exists a scalar $\delta > 0$ such that

$$\|x-y\| \geq \delta \quad \text{whenever} \quad x \in \mathcal{X}^*, y \in \mathcal{X}^*, \psi(x) \neq \psi(y).$$

Note that Proposition 4.12(b) shows that the inequality in part (a) of the assumption holds for strongly convex functions, where the gradient $\nabla f$ is Lipschitz continuous (with $\varepsilon = +\infty$ and $\tau$ being independent of $\zeta$).

Similar assumptions to (a) have been investigated by Luo and Tseng in [105–107]. Note that if a function satisfies the above error bound condition, then it also satisfies the Kurdyka-Łojasiewicz property [98]. Error bounds of this type have been studied by many authors, see e.g. [167, 170].

Some examples of problem classes of the form (1.1) that satisfy Assumption 5.11(a) are the following, cf. [156, 167] and the references therein:

- The function $f$ is strongly convex, $\nabla f$ is Lipschitz continous and $\varphi$ is an arbitrary convex function.
- $f(x) = h(Ax) + c^T x$, where $h : \mathbb{R}^m \to \mathbb{R}$ is a continuously differentiable and strongly convex function such that $\nabla h$ is Lipschitz continuous on every compact set, $A \in \mathbb{R}^{m\times n}, c \in \mathbb{R}^n$, and $\varphi$ has a polyhedral epigraph.
- $f(x) = h(Ax)$, where $A \in \mathbb{R}^{m\times n}$ and $h$ is given as above, and $\varphi(x) = \sum_{i=1}^s \|x_{[\mathcal{I}_i]}\|_2$, where the sets $\mathcal{I}_i \subset \{1, \dots, n\}$ form a partition of $\{1, \dots, n\}$ .

Many more functions of type (1.1) fulfill Assumption 5.11(a) even if they are not covered by the above problem classes. To the author's knowledge, the focus of investigating error bound conditions of these types in literature lies in convex problems. Although not all authors considering the error bound property require convexity of the function $f$, e.g. [156], concrete examples of nonconvex problem classes are very little investigated. For that reason we give an example for a nonconvex function $f$ such that $\psi$ fulfills the Assumption 5.11(a) and note that this might be an interesting topic for future research.

**Example 5.12.** Let $f(x) := \log(1+x^2)$, which is the nonconvex Student's $t$-error function, see [4], and $\varphi(x) := |x|$. The unique minimizer and stationary point of $\psi = f + \varphi$ is $x^* = 0$, so the solution set is $\mathcal{X}^* = \{0\}$. Obviously, this yields $\operatorname{dist}(x, \mathcal{X}^*) = |x|$ and a short calculation shows

$$r(x) = \begin{cases} -1 - \frac{2x}{1+x^2}, & \text{for } x > \alpha, \\ -x & \text{for } |x| \leq \alpha, \\ 1 - \frac{2x}{1+x^2}, & \text{for } x < -\alpha, \end{cases}$$

where $\alpha$ is the real zero of $\alpha^3 - \alpha^2 - \alpha - 1 = 0$.
Let $\zeta \leq \psi(\alpha)$. Then Assumption 5.11(a) holds with $\tau = 1$ for all $x \in \mathbb{R}$ such that $\psi(x) \leq \zeta$. If $\zeta > \psi(\alpha)$, the assumption holds with

$$\tau = \max_{x:\psi(x)\leq\zeta} \left| \frac{x}{1 - 2x/(1+x^2)} \right| > 1.$$

Note that this maximum exists, since $\psi$ is coercive and hence the level set $\operatorname{lev}_{\leq\zeta} \psi$ is compact. Thus, the nonconvex function $\psi$ satisfies the error bound property in Assumption 5.11(a). ◊

For more information and properties of error bound conditions we refer to [156, 167, 170].
Assumption 5.11(b) guarantees that the sets of stationary points of $\psi$ with different function values are properly separated. This assumption holds in particular, if $\psi$ is convex.
It is important to note that we do not assume the convergence of the sequence $\{x^k\}$. Instead, this is a consequence of the above assumptions and the main result in the current section. This theorem is a simplified and slightly adapted version of [156, Theorem 2]. However, for the sake of completeness, we provide the details of the proof.

**Theorem 5.13.** *Let $\{x^k\}$ be a sequence generated by Algorithm 5.1 such that $\nabla f$ is Lipschitz continuous, $MI \succeq H_k \succeq mI$ for some $M \geq m > 0$, and let Assumption 5.11 hold. Then the sequence $\{x^k\}$ is convergent and summable, i.e. $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| < \infty$.*

*Proof.* Since $\{\psi(x^k)\}$ is a nonincreasing sequence, this implies $\psi(x^k) \leq \psi(x^0)$ for all $k \geq 0$. Let $\tau, \varepsilon > 0$ be the constants such that Assumption 5.11(a) holds with $\zeta := \psi(x^0)$. Using Theorem 5.7 and the continuity of $r$, there exists $k_0 \geq 0$ such that $\|r(x^k)\| \leq \varepsilon$ for all $k \geq k_0$. Thus, we get $\|x^k - \overline{x}^k\| \leq \tau \|r(x^k)\|$, where $\overline{x}^k \in \mathcal{X}^*$ satisfies $\|x^k - \overline{x}^k\| = \operatorname{dist}(x^k, \mathcal{X}^*)$ (note that $\overline{x}^k$ does not need to be unique). In particular, since the limit point of the sequence $\{x^k\}$ lies in $\mathcal{X}^*$, this implies $\|x^k - \overline{x}^k\| \to 0$.
By Lemma 5.8(c) we have $\|d^k\| \leq \frac{1+M}{m} \|r(x^k)\|$, which yields $\|x^{k+1} - x^k\| \leq \|d^k\| \to 0$. This and Assumption 5.11(b) imply that the sequence $\{\overline{x}^k\}$ eventually settles down at some isocost surface of $\psi$, i.e. there is $k_1 \geq k_0$ and $\psi^* \in \mathbb{R}$ such that $\psi(\overline{x}^k) = \psi^*$ for all $k \geq k_1$. Fix such an index $k$. Since $\overline{x}^k$ is a stationary point of $\psi$, we have

$$\nabla f(\overline{x}^k)^T (x^k - \overline{x}^k) + \varphi(x^k) - \varphi(\overline{x}^k) \geq 0,$$

as $-\nabla f(\overline{x}^k) \in \partial\varphi(\overline{x}^k)$. In addition, the mean value theorem yields

$$f(x^k) - f(\overline{x}^k) = \nabla f(\xi^k)^T (x^k - \overline{x}^k)$$

for some $\xi^k$ on the line segment between $x^k$ and $\overline{x}^k$. Combining these relations results in

$$\psi^* - \psi(x^k) \leq \big(\nabla f(x^k) - \nabla f(\xi^k)\big)^T (x^k - \overline{x}^k) \leq \big\|\nabla f(x^k) - \nabla f(\xi^k)\big\| \cdot \|x^k - \overline{x}^k\| \leq L\|x^k - \overline{x}^k\|^2.$$

This, together with $\{x^k - \overline{x}^k\} \to 0$ proves that

$$\liminf_{k\to\infty} \psi(x^k) \geq \psi^*. \tag{5.7}$$

By definition of $d^k$, applying Fermat's rule twice yields

$$d^k \in \arg\min_{d\in\mathbb{R}^n} \big\{(\nabla f(x^k) + H_k d^k)^T d + \varphi(x^k + d)\big\}.$$

In particular, we have

$$\big(\nabla f(x^k) + H_k d^k\big)^T d^k + \varphi(x^k + d^k) \leq \big(\nabla f(x^k) + H_k d^k\big)^T (\overline{x}^k - x^k) + \varphi(\overline{x}^k),$$

which is equivalent to

$$\big(\nabla f(x^k) + H_k d^k\big)^T (x^k + d^k - \overline{x}^k) + \varphi(x^k + d^k) - \varphi(\overline{x}^k) \leq 0. \tag{5.8}$$

Let $\mathcal{S} \subset \mathbb{N}_0$ be the set of all successful or highly successful steps in Algorithm 5.1, and fix $k \in \mathcal{S}$ with $k \geq k_1$ for the moment. Now, we use the mean value theorem again for some $\tilde{\xi}^k$ on the straight line between $x^{k+1}$ and $\overline{x}^k$ to obtain

$$\begin{aligned}
\psi(x^{k+1}) - \psi^* &= f(x^{k+1}) + \varphi(x^{k+1}) - f(\overline{x}^k) - \varphi(\overline{x}^k) \\
&= \nabla f(\tilde{\xi}^k)^T (x^{k+1} - \overline{x}^k) + \varphi(x^{k+1}) - \varphi(\overline{x}^k) \\
&= \big(\nabla f(\tilde{\xi}^k) - \nabla f(x^k)\big)^T (x^{k+1} - \overline{x}^k) - (H_k d^k)^T (x^{k+1} - \overline{x}^k) \\
&\qquad + \big[\big(\nabla f(x^k) + H_k d^k\big)^T (x^k + d^k - \overline{x}^k) + \varphi(x^{k+1}) - \varphi(\overline{x}^k)\big] \\
&\leq L\|\tilde{\xi}^k - x^k\| \cdot \|x^{k+1} - \overline{x}^k\| + M\|d^k\| \cdot \|x^{k+1} - \overline{x}^k\|,
\end{aligned} \tag{5.9}$$

where we used (5.8) for the final estimate. By assumption, we have $(d^k)^T(H_k + \mu_k I)d^k \geq m\|d^k\|^2$. Furthermore, we use the error bound condition, Lemma 5.8(b), and $\mu_k \leq \mu_{\max}$ for some $\mu_{\max} > 0$ (Proposition 5.10) to obtain

$$\|x^k - \overline{x}^k\| \leq \tau\|r(x^k)\| \leq \tau\frac{m+1}{m}(M + \mu_{\max})\|d^k\| =: \tilde{C}\|d^k\|.$$

Using, in addition, the estimates

$$\|\tilde{\xi}^k - x^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \overline{x}^k\|, \quad \text{and} \quad \|x^{k+1} - \overline{x}^k\| \leq \|x^{k+1} - x^k\| + \|x^k - \overline{x}^k\|,$$

we get from (5.9)

$$\psi(x^{k+1}) - \psi^* \leq C\|d^k\|^2 \qquad \text{with } C = L(1 + \tilde{C})^2 + M(1 + \tilde{C}) > 0.$$

Using again Lemma 5.8(a), for $k \in \mathcal{S}$ with $k \geq k_1$ there holds

$$-\operatorname{ared}_k = \psi(x^{k+1}) - \psi(x^k) \leq -c_1 \operatorname{pred}_k \leq -\frac{1}{2}c_1 m\|d^k\|^2. \tag{5.10}$$

Thus, with (5.7), we obtain

$$0 \le \psi(x^{k+1}) - \psi^* \le C\|d^k\|^2 \le C'\big(\psi(x^k) - \psi(x^{k+1})\big)$$

for $C' := 2C/(c_1 m)$, which is equivalent to

$$\psi(x^{k+1}) - \psi^* \le \frac{C'}{1+C'}\big(\psi(x^k) - \psi^*\big).$$

Hence, $\{\psi(x^k)\}_{\mathcal{S}}$, converges to $\psi^*$ at least linearly. Finally, using again $\mathrm{ared}_k \ge c_1 \,\mathrm{pred}_k \ge \frac{c_1 m}{2}\|d^k\|^2$, we get $\psi(x^k) - \psi(x^{k+1}) \ge c_1 m/2\|d^k\|^2$, which is equivalent to

$$\|d^k\| = \|x^{k+1} - x^k\| \le \sqrt{\frac{2}{c_1 m}\big(\psi(x^k) - \psi(x^{k+1})\big)} \le \sqrt{\frac{2}{c_1 m}\big(\psi(x^k) - \psi(x^*)\big)} \le C''\vartheta^{s(k)}$$

for some $C'' > 0$ and $\vartheta := \sqrt{\frac{C'}{1+C'}}$, where $s(k) \in \mathbb{N}$ denotes the number of elements in $\mathcal{S}$ less than or equal to $k$. Thus,

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| = \sum_{k\in\mathcal{S}} \|x^{k+1} - x^k\| \le C'' \sum_{k\in\mathcal{S}} \vartheta^{s(k)} = C'' \sum_{k=0}^{\infty} \vartheta^k = \frac{C''}{1-\vartheta} < +\infty,$$

$\{x^k\}$ is a Cauchy sequence and therefore convergent to some $x^*$, which completes the proof. □

## 5.4 Local Convergence and Convergence Rates

The aim of this section is to develop the local convergence theory for Algorithm 5.1. In addition to the assumptions of the previous section (Lipschitz continuity of $\nabla f$, uniform boundedness and positive definiteness of the sequence $\{H_k\}$ and the local error bound from Assumption 5.11), we assume that $f$ is twice continuously differentiable and that the Dennis-Moré condition is satisfied.

More precisely, let $\{x^k\}$ be a sequence generated by Algorithm 5.1 under the above assumptions. By Theorem 5.13 this sequence converges to a stationary point $x^* \in \mathbb{R}^n$ of $\psi$. The corresponding sequences $\{H_k\}$ and $\{d^k\}$ are then supposed to satisfy the Dennis-Moré condition

$$\lim_{k\to\infty} \frac{\|(H_k - \nabla^2 f(x^*))d^k\|}{\|d^k\|} = 0, \tag{5.11}$$

analogous to Section 4.3, cf. [56]. Note that with Theorem 5.13 and the arguments used in Remark 4.6, it follows that this property is satisfied for Algorithm 5.1, if $\{H_k\}$ is updated by the BFGS-scheme (3.15).

Under these conditions, we prove our final convergence theorem. The next step towards this result is an estimate on the ratio of $\mathrm{ared}_k$ and $\mathrm{pred}_k$.

**Lemma 5.14.** *Assume that $\nabla f$ is Lipschitz continuous with Lipschitz constant $L > 0$, $MI \succeq H_k \succeq mI$ holds for some $M \ge m > 0$, $\{H_k\}$ satisfies* (5.11)*, and Assumption 5.11 holds. Let the sequence $\{x^k\}$ be generated from Algorithm 5.1 and let $x^*$ be its limit point. Then, for any $c \in (0,1)$ there exists $\varepsilon > 0$ (depending on $c$) such that*

$$\mathrm{ared}_k - c\,\mathrm{pred}_k > 0$$

*holds for all $x^k \in B_\varepsilon(x^*)$.*

*Proof.* Since $f$ is twice continuously differentiable, for every $k \geq 0$ there exists $\xi^k$ on the line segment between $x^k$ and $x^k + d^k$ such that

$$f(x^k + d^k) - f(x^k) - \nabla f(x^k)^T d^k = \frac{1}{2}(d^k)^T \nabla^2 f(\xi^k) d^k.$$

Then

$$\begin{aligned}
&\psi(x^k + d^k) - \psi(x^k) + q_k(0) - q_k(d^k) \\
&\quad = f(x^k + d^k) - f(x^k) - \nabla f(x^k)^T d^k - \frac{1}{2}(d^k)^T H_k d^k \\
&\quad = \frac{1}{2}(d^k)^T \big(\nabla^2 f(\xi^k) - H_k\big) d^k \\
&\quad \leq \frac{1}{2}\big\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\big\| \, \|d^k\|^2 + \frac{1}{2}\big\|\nabla^2 f(x^k) - \nabla^2 f(x^*)\big\| \, \|d^k\|^2 \\
&\qquad + \frac{1}{2}\big\|(H_k - \nabla^2 f(x^*))d^k\big\| \, \|d^k\|. \qquad (5.12)
\end{aligned}$$

Note that Lemma 5.8(c) and Proposition 4.12(a) yield

$$\|d^k\| \leq \frac{1+M}{m}(2+L)\,\mathrm{dist}(x^k, \mathcal{X}^*) \leq \frac{1+M}{m}(2+L)\|x^k - x^*\| \leq \frac{1+M}{m}(2+L)\varepsilon$$

if $x^k \in B_\varepsilon(x^*)$, where $\mathcal{X}^*$ denotes the set of stationary points of $\psi$. Due to the continuity of $\nabla^2 f$ and (5.11) we can choose $\varepsilon > 0$ sufficiently small such that the following estimates hold:

$$\begin{aligned}
\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| &\leq \frac{m}{6}(1-c), \\
\|\nabla^2 f(x^k) - \nabla^2 f(x^*)\| &\leq \frac{m}{6}(1-c), \\
\|(H_k - \nabla^2 f(x^*))d^k\| &\leq \frac{m}{6}(1-c)\|d^k\|,
\end{aligned}$$

whenever $\|x^k - x^*\| \leq \varepsilon$. Inserting these inequalities in (5.12) yields

$$\psi(x^k + d^k) - \psi(x^k) + q_k(0) - q_k(d^k) \leq \frac{m}{4}(1-c)\|d^k\|^2.$$

Furthermore, by Lemma 5.8(a) we have $q_k(0) - q_k(d^k) = \mathrm{pred}_k \geq \frac{m}{2}\|d^k\|^2$. The combination of both estimates results in

$$\begin{aligned}
\mathrm{ared}_k - c\,\mathrm{pred}_k &= \psi(x^k) - \psi(x^k + d^k) - c\big(q_k(0) - q_k(d^k)\big) \\
&= (1-c)\big(q_k(0) - q_k(d^k)\big) - \big(\psi(x^k + d^k) - \psi(x^k) + q_k(0) - q_k(d^k)\big) \\
&\geq \frac{m}{4}(1-c)\|d^k\|^2.
\end{aligned}$$

The claim follows, since, by assumption, there holds $d^k \neq 0$ (otherwise $x^k$ is already a stationary point of $\psi$). □

In the next result we summarize the convergence results, which hold under the Dennis-Moré-condition. We note that part (a), which states that finally all iterations are highly successful, can also be formulated for successful (not necessarily highly successful) steps, with possibly smaller value for $k_0 \geq 0$.

**Theorem 5.15.** *Consider a sequence $\{x^k\}$ generated by Algorithm 5.1 and let the assumptions of Lemma 5.14 hold. Then the following hold:*

*(a) There exists $k_0 \geq 0$ such that all steps $k \geq k_0$ are highly successful.*

*(b) The sequence $\{\mu_k\}$ converges to zero.*

*(c) The sequence $\{\operatorname{dist}(x^k, \mathcal{X}^*)\}$ converges to zero superlinearly, where $\mathcal{X}^*$ is the set of stationary points of $\psi$.*

*Proof.* Assuming in contradiction to the claim of part (a) that there are infinitely many unsuccessful iterates, similar to the proof of Proposition 5.10 in combination with Lemma 5.14 we get the estimate (5.6)

$$\left(1 + \frac{1}{m + \mu_k}\right) \frac{M + \mu_k}{\mu_k} \geq \frac{m + 2\mu_k}{2p_{\min}\mu_k} \|r(x^k)\|^{1-\kappa},$$

for unsuccessful steps $k$, which is equivalent to

$$\|r(x^k)\|^{\kappa-1} \geq \frac{1}{2p_{\min}} \cdot \frac{m + 2\mu_k}{M + \mu_k} \cdot \frac{m + \mu_k}{1 + m + \mu_k} \geq \frac{1}{2p_{\min}} \cdot \frac{m^2}{M(1+m)}.$$

Hence, noting that $\kappa > 1$, there is a lower bound for $\|r(x^k)\|$ restricted to unsuccessful iterates, which is a contradiction to Theorem 5.7. This proves part (a). Part (b) is a direct consequence of (a) since $\mu_k$ is multiplied by $\sigma_1 < 1$ for all highly successful steps.

It remains to prove (c). Let $k \geq k_0$. Then $x^{k+1} = \arg\min_x \hat{q}_k(x - x^k)$. We write $\hat{q}_k(d) =: f_k(d) + \varphi(x^k + d)$, where $f_k$ is a smooth function, and adapt the characterization of stationarity (Corollary 3.8) to this function to obtain

$$x^{k+1} = \operatorname{prox}_\varphi\left(x^{k+1} - \nabla f_k(d^k)\right) = \operatorname{prox}_\varphi\left(x^{k+1} - \nabla f(x^k) - (H_k + \mu_k I)(x^{k+1} - x^k)\right).$$

By the error bound property (Assumption 5.11(a)) there exist $\varepsilon, \tau > 0$, depending on $x^0$, such that

$$\operatorname{dist}(x, \mathcal{X}^*) \leq \tau \cdot \|r(x)\|, \qquad \text{whenever } \psi(x) \leq \psi(x^0), \|r(x)\| \leq \varepsilon.$$

By Theorem 5.7 we have $\lim_{k\to\infty} \|r(x^k)\| = 0$. Thus, $\|r(x^k)\| \leq \varepsilon$ holds for all $k \geq k_0$, by possibly enlarging $k_0$. Furthermore, Algorithm 5.1 ensures that $\psi(x^{k+1}) \leq \psi(x^k)$. Hence, we get $\operatorname{dist}(x^k, \mathcal{X}^*) \leq \tau\|r(x^k)\|$ for all $k \geq k_0$. Applying this property to $x^{k+1} = x^k + d^k$, we obtain

$$\begin{aligned}
&\frac{1}{\tau}\operatorname{dist}(x^{k+1}, \mathcal{X}^*) \leq \|r(x^{k+1})\| \\
&\quad = \left\| \operatorname{prox}_\varphi(x^{k+1} - \nabla f(x^{k+1})) - x^{k+1} \right\| \\
&\quad = \left\| \operatorname{prox}_\varphi(x^{k+1} - \nabla f(x^{k+1})) - \operatorname{prox}_\varphi\left(x^{k+1} - \nabla f(x^k) - (H_k + \mu_k I)(x^{k+1} - x^k)\right) \right\| \\
&\quad \leq \left\| \nabla f(x^{k+1}) - \nabla f(x^k) - (H_k + \mu_k I)(x^{k+1} - x^k) \right\| \\
&\quad \leq \left\| \nabla f(x^{k+1}) - \nabla f(x^k) - H_k d^k \right\| + \mu_k \|d^k\| \\
&\quad \leq \left\| \nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k) d^k \right\| + \left\| \left(\nabla^2 f(x^k) - \nabla^2 f(x^*)\right) d^k \right\| \\
&\qquad + \left\| \left(\nabla^2 f(x^*) - H_k\right) d^k \right\| + \mu_k \|d^k\|
\end{aligned}$$

where the inequality in the forth line follows from the nonexpansivity of the proximity operator (Proposition 3.5). As $f$ is twice continuously differentiable, the first two terms are $o(\|d^k\|)$ for $\|d^k\| \to 0$. The third term is $o(\|d^k\|)$ by the Dennis-Moré-condition and

$\mu_k \|d^k\| = o(\|d^k\|)$ follows from $\mu_k \to 0$. Hence, $\operatorname{dist}(x^{k+1}, \mathcal{X}^*) \leq o(\|d^k\|)$. The claim follows using Lemma 5.8(c) and the estimate $\|r(x^k)\| \leq (2+L)\operatorname{dist}(x^k, \mathcal{X}^*)$, which is shown in the same way as in the proof of Proposition 4.12(a). □

**Remark 5.16.** The approach of the local convergence theory in this chapter differs from the one in Chapter 4. There, the results are deduced under the assumption that an accumulation point exists and the preliminaries hold in an appropriate neighbourhood of that point. In contrast, the convergence theory in Sections 5.3 and 5.4 has some restricting (global) assumptions as global Lipschitz continuity of $\nabla f$, but shows that in that case the entire sequence of iterates is convergent.
With some adaptations of this approach, however, it is possible to reformulate the theory of this chapter in analogy to the previous one. In particular, assuming that there exists an accumulation point of the sequence of iterates, the stated preliminaries only need to hold in an appropriate neighbourhood of this point. ♢

## 5.5 A Modified Version using Proximal Gradient Framework

A drawback of the regularized proximal quasi-Newton method from Algorithm 5.1 is that in unsuccessful iterations $x^k$ is not updated and the algorithm remains at this point until a successful or highly successful iteration is performed. The relevant consequence of this procedure is that $x^k$ and $H_k$ might be such that multiple unsuccessful steps follow, and therefore multiple quasi-Newton subproblems must be solved without benefit. To circumvent this issue, in this section we present a modified method whose idea is to update the iterate using a proximal gradient-like approach in the case of unsuccessful steps. We note that there is no advantage over Algorithm 5.1 concerning the convergence theory, since Algorithm 5.1 is already globally convergent. However, experiments show that the numerical behaviour can be improved for both, convex and nonconvex problem settings. First, we deduce the modification and then investigate its convergence properties.

In Algorithm 5.1 the solution of the subproblem (5.1) is discarded in unsuccessful steps. Since this search direction $d^k$ is the solution of a (possibly computationally expensive) subproblem, one idea might be to use the search direction $d^k$ anyway, in case it is a direction of descent, and combine it with some line search strategy, e.g. of Armijo-type similar to (4.8). This would ensure that the particular proximal quasi-Newton search direction is used as many times as possible. On the other hand, this search direction is not too good, since the corresponding step would otherwise be (highly) successful. Therefore, in case of unsuccessful steps, it may be more natural to determine a search direction as the solution of a proximal gradient-type subproblem instead (similar to Algorithm 4.1), i.e. to set $d^k = r_{\tau_k I}(x^k)$ for a suitable $\tau_k > 0$. To obtain the next iterate, an Armijo-type line search is performed. For this purpose, set

$$\Delta_k := \nabla f(x^k)^T d^k + \varphi(x^k + d^k) - \varphi(x^k).$$

If a step is successful or highly successful, the approach of the method coincides with the regularized proximal quasi-Newton method in Algorithm 5.1. The complete method using this modification is presented in Algorithm 5.2.

To differentiate from the unsuccessful steps in Algorithm 5.1, where the iterate is not updated, note that the iterations in Algorithm 5.2, in which (S.4) is performed, are called *semi-successful.* Similar to Algorithm 5.1 the regularization parameter $\mu_k$ in Algorithm

**Algorithm 5.2** Modified Regularized Proximal Quasi-Newton Method

```
(S.0) Choose $x^0 \in \mathbb{R}^n$, parameters $\mu_0 > 0$, $p_{\min} \in (0,1/2)$, $\kappa > 1$, $c_1 \in (0,1/2)$,
      $c_2 \in (c_1,1)$, $\sigma_1 \in (0,1), \sigma_2 > 1$, $0 < \tau_{\min} \leq \tau_{\max}$, $\beta, \sigma \in (0,1)$ and set $k := 0$.
(S.1) If $x^k$ satisfies a suitable termination criterion: STOP.
(S.2) Choose $H_k \in \mathbb{R}^{n\times n}$, and find a solution $d^k$ of the problem
          $\min_d \hat{q}_k(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T (H_k + \mu_k I) d + \varphi(x^k + d).$
      If this problem has no solution, or if
          $\mathrm{pred}_k \leq p_{\min} \|d^k\| \cdot \min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\},$
      go to (S.4). Otherwise go to (S.3).
(S.3) Set $\rho_k := \mathrm{ared}_k \,/\, \mathrm{pred}_k$ and perform the following updates:
      • If $\rho_k > c_2$ (highly successful step), set $x^{k+1} = x^k + d^k$ and $\mu_{k+1} = \sigma_1 \mu_k$,
      • If $c_2 \geq \rho_k > c_1$ (successful step), set $x^{k+1} = x^k + d^k$ and $\mu_{k+1} = \mu_k$,
      and go to (S.5). If $\rho_k \leq c_1$, go to (S.4).
(S.4) Choose $\tau_k \in [\tau_{\min}, \tau_{\max}]$, set $d^k = r_{\tau_k I}(x^k)$, compute $t_k = \max\{\beta^l : l = 0,1,2,\dots\}$
      such that
          $\psi(x^k + t_k d^k) \leq \psi(x^k) + \sigma t_k \Delta_k,$
      set $x^{k+1} = x^k + t_k d^k$ and $\mu_{k+1} = \sigma_2 \mu_k$ (semi-successful step).
(S.5) Update $k \leftarrow k + 1$, and go to (S.1).
```

5.2 is increased in semi-successful steps, since our aim is to preferably perform (highly) successful steps.

We note that Algorithm 5.2 is well-defined and there holds $\psi(x^{k+1}) \leq \psi(x^k)$ for all $k \geq 0$. For successful and highly successful steps this follows from $\rho_k > 0$ and therefore $\mathrm{ared}_k = \psi(x^k) - \psi(x^k + d^k) > 0$; for semi-successful steps this follows similar to Proposition 3.10.

**Remark 5.17.** For semi-successful steps, a new search direction is determined in Algorithm 5.2 and the search direction computed in (S. 2) is discarded. As mentioned at the beginning of this section, the search direction determined in (S.2) can also be used in (S.4) instead of the proximal gradient update, as long as it is at least a descent direction. The new iterate is then determined by some line search. To check whether we have a descent direction, for example a criterion like (4.7) in Algorithm 4.1 can be used. Moreover, we have to investigate the case that $d^k$ is not a descent direction. Two possibilities are obvious for handling this issue: Either no update of $x^k$ is performed as in unsuccessful steps in Algorithm 5.1, or an update follows as in Algorithm 5.2. The global convergence theorem differs in this case because, although $\|d^k\| \to 0$ can be shown analogous to Theorem 4.3, it is not possible to prove $\|r(x^k)\| \to 0$ without further assumptions. Since this variant was less convincing in numerical test runs than the one presented in Algorithm 5.2, the former will not be discussed further in the following. ◊

We briefly discuss the convergence properties of the modified version of the regularized proximal quasi-Newton method. It is worth noting that the modification has little effect on the theory already shown for Algorithm 5.1 in the previous sections. We start with the global convergence result analogous to Theorem 5.7.

**Theorem 5.18.** *Let $\{H_k\}$ be a bounded sequence of symmetric matrices, assume that $\psi$ is bounded from below and that $\nabla f$ is Lipschitz continuous on a set $\mathcal{X}$ satisfying $\{x^k\} \subset \mathcal{X}$, where $\{x^k\}$ denotes a sequence generated by Algorithm 5.2. Then $\lim_{k\to\infty} \|r(x^k)\| = 0$ holds; in particular, every accumulation point of $\{x^k\}$ is a stationary point of $\psi$.*

*Proof.* Set

$$\mathcal{K}_S := \{k : d^k \text{ is obtained in (S.2)}\}, \qquad \text{and} \qquad \mathcal{K}_G := \{k : d^k \text{ is obtained in (S.4)}\}.$$

Hence, $\mathcal{K}_S$ is the set of all successful and highly successful iterations, and $\mathcal{K}_G$ is the set of semi-successful iterations. If $\mathcal{K}_G$ is finite, the claim follows from Theorem 5.7. So, assume that there are infinitely many semi-successful steps and let $k \in \mathcal{K}_G$. The line search in (S.4), Lemma 3.9 and the lower bound $t_{\min} > 0$ for the step size $t_k$ (Lemma 3.13) yield

$$\psi(x^{k+1}) - \psi(x^k) \le \sigma t_k \Delta_k \le -\sigma t_{\min} \tau_{\min} \|d^k\|^2 < 0. \tag{5.13}$$

Thus, using the boundedness of $\psi$ from below, $\lim_{k\in\mathcal{K}_G} \|d^k\| = 0$ and Lemma 3.6 yields $\lim_{k\in\mathcal{K}_G} \|r(x^k)\| = 0$. Assume that there is a subsequence $\mathcal{K} \subset \mathbb{N}_0$ and $\delta > 0$ such that $\|r(x^k)\| \ge 2\delta$ for all $k \in \mathcal{K}$. For any $k \in \mathcal{K}$ let $\ell(k) > k$ be such that $\|r(x^l)\| > \delta$ for all $l = k+1, \dots, \ell(k)-1$ and $\|r(x^{\ell(k)})\| \le \delta$. By the above, we can assume without loss of generality that $\mathcal{K} \subset \mathcal{K}_S$ and $k+1, \dots, \ell(k)-1 \in \mathcal{K}_S$ for all $k \in \mathcal{K}$. Furthermore, set $\overline{\delta} := \min\{\delta, \delta^\kappa\}$. Now, the remaining part of the proof coincides with the one of Theorem 5.7. Since all iterations $k \le l < \ell(k)$ are successful or highly successful for $k \in \mathcal{K}$, we get

$$\psi(x^l) - \psi(x^{l+1}) \ge c_1 \operatorname{pred}_l \ge c_1 p_{\min} \overline{\delta} \|x^{l+1} - x^l\|.$$

Thus,

$$p_{\min} c_1 \overline{\delta} \|x^{\ell(k)} - x^k\| \le p_{\min} c_1 \overline{\delta} \sum_{l=k}^{\ell(k)-1} \|x^{l+1} - x^l\| \le \sum_{l=k}^{\ell(k)-1} \psi(x^l) - \psi(x^{l+1}) = \psi(x^k) - \psi(x^{\ell(k)})$$

holds for all $k \in \mathcal{K}$. By assumption, $\psi$ is bounded from below, and by construction, the sequence $\{\psi(x^k)\}$ is monotonically decreasing, hence convergent. This implies $\{\psi(x^k) - \psi(x^{\ell(k)})\} \to_{\mathcal{K}} 0$. Hence, we get $\{\|x^{\ell(k)} - x^k\|\} \to_{\mathcal{K}} 0$. The Lipschitz continuity of the residual function $r(\cdot)$ yields $\{\|r(x^{\ell(k)}) - r(x^k)\|\} \to_{\mathcal{K}} 0$. On the other hand, by the choice of $\ell(k)$, we have

$$\left\|r(x^k) - r(x^{\ell(k)})\right\| \ge \left\|r(x^k)\right\| - \left\|r(x^{\ell(k)})\right\| \ge 2\delta - \delta \ge \delta,$$

which yields the desired contradiction. Hence, $\lim_{k\to\infty} \|r(x^k)\| = 0$. □

We note that the Lipschitz continuity of $\nabla f$, which is stronger than the uniform continuity assumed in Theorem 5.7 is necessary to ensure that the step size $t_k$ in (S.4) of Algorithm 5.2 is bounded from below.

A closer look at the proofs of the results in Sections 5.3 and 5.4 shows that they can be applied almost in the same way to the modified Algorithm 5.2. In summary, assuming that $\psi$ is bounded from below, $\nabla f$ is Lipschitz continuous, and the sequence $\{H_k\}$ is uniformly bounded and positive definite, i.e. there exist $M \ge m > 0$ such that $MI \succeq H_k \succeq mI$, the following results hold for Algorithm 5.2:

- The sequence $\{\mu_k\}$ is bounded. (Proposition 5.10)

- Under the error bound assumption (Assumption 5.11) the sequence $\{x^k\}$ converges to a stationary point $x^* \in \mathbb{R}^n$ and $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| < +\infty$. (Theorem 5.13)
- Under the error bound assumption (Assumption 5.11) and the Dennis-Moré-condition (5.11) finally all steps are highly successful,
- the sequence $\{\mu_k\}$ converges to 0,
- and the sequence $\{\text{dist}(x^k, \mathcal{X}^*)\}$ converges to 0 superlinearly, where $\mathcal{X}^*$ denotes the set of stationary points of $\psi$. (Theorem 5.15)

A single modification is needed in the proof of the result corresponding to Proposition 5.10: Here, in addition to equation (5.10) for successful and highly successful steps, we need an analogous estimate for semi-successful steps. This results follows from the line search in (S.4) of Algorithm 5.2, Lemma 3.9 and the lower bound for the step size $t_k$ (Lemma 3.13), see (5.13). Details are left to the reader.

# Chapter 6

# Applications and Numerical Results

The purpose of this chapter is to provide an extensive numerical investigation of the methods presented in the previous chapters using numerous examples relevant in applications. Our methods are not only analyzed in terms of their practical performance, but also compared to several state-of-the-art methods with a focus on proximal methods. Let us note that parts of the following sections are essentially based on the work [82] and that several parts have already appeared in a similar form in [81, 82]. However, most of the results reported in this thesis were obtained by using more refined and improved versions of the algorithms. The chapter is organized as follows. In Section 6.1 we start with the description of several state-of-the-art methods, which are used in comparison to our methods and give an overview of parameters that occur within the methods throughout the chapters. Afterwards, we start the investigation of numerical examples. First, Section 6.2 provides a simple, two-dimensional example to make some introductory comments. In Section 6.3 we consider logistic regression problems with $\ell_1$-regularizer. The main purpose of this section is to study the properties of the globalized inexact proximal Newton method of Section 4 and the competitive ability of Algorithm 3.2 for the solution of the subproblems. Section 6.4 examines quadratic problems with the group sparse regularizer. Here, the focus lies on the details and performance of the regularized proximal Newton method introduced in Chapter 5. The following sections deal with nonconvex problems using Student's $t$-regression and regularization by an $\ell_1$-term. In Section 6.5 we use synthetic data, while Section 6.6 covers an example of nonconvex image reconstruction. Finally, the example stated in Section 6.7 investigates proximal methods with regard to a not analytically computable proximity operator.

The numerical results have been obtained in MATLAB R2020b using a machine running Open SuSE Leap 15.2 with an Intel core i5 processor 3.2GHz and 16 GB RAM with the exception of the results in Section 6.7. These come from tests in MATLAB R2018b running Open SuSE Leap 15.1 on the same machine.

## 6.1 State-of-the-art Methods and Implementation

This section consists of two parts. First, in Section 6.1.1, we present details of the algorithms, which are used in the following for the comparison with the methods presented in this thesis. Second, we use Section 6.1.2 to make some comments on the implementation of our algorithms.

### 6.1.1 State-of-the-art Methods

In this section we state main ideas and basic structural aspects of several state-of-the-art methods, which will be used in our numerical comparison. In the subsequent analysis we focus on proximal methods, since these profoundly exploit the composite structure of problem (1.1) and turned out to outperform most methods which make use of different approaches. Let us begin with the details of some first-order proximal methods.

**Fast Iterative Shrinkage Thresholding Algorithm (FISTA)** FISTA by Beck and Teboulle [14] is perhaps the best known and most widely used algorithm for the solution of convex composite optimization problems of the form (1.1). It is a first-order proximal method with acceleration based on the work of Nesterov [120] for problems with convex $f$ such that $\nabla f$ is (globally) Lipschitz continuous. In every step a problem of type (1.2) is solved for $H_k = L_k I$, where $L_k$ is an approximation to the Lipschitz constant of $\nabla f$, which is repeatedly increased using a backtracking strategy until a sufficient decrease condition holds. Although there exist a couple of adaptations of FISTA for nonconvex problems [99, 129], we decided to implement only the original version of this algorithm.
For the approximation of the Lipschitz constant we initialize the method with $L_0 = 1$ and use the increasing factor $\eta = 2$.

**Sparse Reconstruction by Separable Approximation (SpaRSA)** The method developed by Wright et al. [164] is another accelerated first order proximal method. The main difference to FISTA is the update of the factor $L_k$, which is done using a Barzilai-Borwein approach. Furthermore, the acceptance criterion

$$\psi(x^{k+1}) \leq \max_{i=\max\{k-M,0\},\ldots,k} \psi(x^i) - \frac{\sigma}{2}\alpha_k \|x^{k+1} - x^k\|^2$$

in each step is nonmonotone. As a consequence, the convergence theory covers both, convex and nonconvex problems with Lipschitz continuous $\nabla f$ and a real-valued convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$.
With the notation of [164] the initial value for the proximity scaling is set to $\alpha_0 = 1$, whereas $\alpha_{\min} = 10^{-4}$ and $\alpha_{\max} = 10^4$. If the acceptance criterion is not met, the parameter is increased with $\eta = 2$. Furthermore, as suggested in [164], we use $M = 5$ and $\sigma = 0.01$. In contrast to their approach we do not apply an adaptive continuation strategy in order to get the basic versions of all methods.

**Proximal Gradient Method (PG)** The elaboration of the proximal gradient method with Armijo-type line search was given in Section 3.2. Despite the exhaustive presentation in this thesis, this first-order proximal method is not of prime importance in our numerical analysis and, hence, listed as state-of-the-art method.
Maybe most important in the implementation of the method is the choice of the matrices $H_k = \tau_k I$. The parameter $\tau_k \in [\tau_{\min}, \tau_{\max}]$ is successively computed to approximate a (local) Lipschitz constant of $\nabla f$. For that purpose, in each step we compute $\hat{\tau}_k = \|\nabla f(x^k) - \nabla f(x^{k-1})\| / \|x^k - x^{k-1}\|$ as an approximation in the current step and set $\tau_k$ to be a weighted mean of $\tau_{k-1}$ and $\hat{\tau}_k$ similar to the approach in [99].
In addition, we use $\tau_{\min} = 10^{-4}$, $\tau_{\max} = 10^4$, $\beta = 0.1$ and $\sigma = 10^{-4}$.

**Semismooth Newton Method with Filter Globalization (SNF)** This second-order method by Milzarek and Ulbrich [111, 112] is based on the semismooth Newton method to

find a zero of $r(x)$ defined in (4.4), combined with a globalization using a filter strategy. They provide a convex and nonconvex version of the filter conditions to decide whether the computed update is applied or a proximal gradient step is performed instead.
The filter and the parameters are chosen based on the suggestions in [111, 112]. In detail, the parameters for the acceptance condition in the filter update are $\alpha_1 = \alpha_2 = \alpha_3 = 10^{-1}$, $\eta = 0.8$ and the constants for computing the proximal gradient steps coincide with the ones described for PG.

A drawback of this method is that it can only be applied, when $r(x)$ and its Newton derivative can be computed, hence, not in the example in Section 6.7. For this instance we provide one more method.

**Primal-Dual Fixed Point Method based on the Proximity Operator (PDFP$^2$O)**
The fixed point method by Chen et al. [43] reformulates problem (1.1) to get a nonsmooth fixed point equation, in the case that the nonsmooth function $\varphi$ is replaced by $\varphi \circ B$ for a linear transformation $B : \mathbb{R}^n \to \mathbb{R}^m$. For the convergence theory, $f$ must be continuous and the Lipschitz constant of the Lipschitz continuous gradient $\nabla f$ must be known. Thus, the applicability of this method is limited, but we will use it in Section 6.7.
Parameters are set as suggested by the authors in [43].

### 6.1.2 Details on the Implementation

In this section we give several details on the implementation of the (inexact) globalized proximal Newton method (GPN) and an inexact globalized proximal quasi-Newton method (QGPN) as described in Algorithm 4.1, as well as the regularized proximal quasi-Newton method (RPQN) of Algorithm 5.1. Although the focus of Chapter 5 was on the quasi-Newton formulation, we refer to the method as regularized proximal Newton method (RPN), when using the exact Hessian of the smooth function, while the modified method in Algorithm 5.2 is abbreviated by RPQNm or RPNm, respectively.

**Subproblem Solvers** The crucial part of the implementation of the mentioned algorithms is the efficient solution of the subproblems (1.2). In the case that the matrices $H_k$ are updated by means of a quasi-Newton approach, Algorithmus 3.2 is applied. This technique will be addressed in the next but one paragraph.
However, especially when running the versions GPN and RPN that make use of the exact Hessian, we need to apply different solvers for the subproblems. As described in [82], two methods are implemented for this purpose: FISTA and SNF. The details on these methods are given in the previous section. To point out which subproblem solver is currently used, we add the particular first letter to the abbreviations, i.e. GPN-F and GPN-S for the globalized proximal Newton method with subproblem solver FISTA and SNF, respectively. Experiments show that the methods perform best, when we stop after a maximum of 80 iterations for FISTA or 10 iterations for SNF, unless the termination criterion is satisfied in earlier iterations.
More details and discussion regarding the solution of the subproblems is provided in Section 6.3.3.

**Inexactness** The theory of Algorithm 4.1 covers the possibility for solving the subproblems inexactly. To address this fact, instead of implementing criterion (4.6) directly, Algorithm 4.1 turned out to perform better if we stop the subproblem solvers FISTA or SNF with a low maximal number of iterations as explained in the previous paragraph. Although the

theory of the regularized proximal quasi-Newton method does not come together with some inexactness strategy, we also terminate RPN-S after at most 10 iterations.
However, a detailed investigation of the inexact solution of the subproblems and the inexactness criterion in (4.6) is given in Section 6.3.2.

**Quasi-Newton Approach** A major strength of the presented second-order methods is the high efficiency especially in combination with a suitable quasi-Newton strategy for updating the matrices $H_k$. In that case, primarily Algorithm 3.2, which has already proven to be very precise after one or two iterations in numerical test runs, is used to solve the subproblems. The maximum iteration number is therefore set to 10 and we do not apply inexactness strategies for these methods. In detail, we use $\eta = 0$ in Algorithm 4.1, and stop Algorithm 3.2 if $\|\mathcal{L}(\alpha)\| < 10^{-10}$. A comparison to other solvers for the subproblems is given in Section 6.3.3. Note that the method to solve the subproblems in Algorithm 3.2 also applies to the regularized matrix $H_k + \mu_k I$, because the form (3.17) is preserved by adding $\mu_k I$ to the simple matrix $H_0$.
Assuming that $\psi$ satisfies the preliminaries of the convergence theorems in Sections 4 or 5, a sequence $\{H_k\}$ generated using BFGS-updates satisfies the Dennis-Moré-condition, cf. Remark 4.6. Motivated by this idea and with the discussion in Section 3.3 we implement the algorithms QGPN and RPQN, with $H_k$ being a limited memory BFGS- or SR1-update, denoted by L-BFGS or L-SR1 in the following. The size of the memory is chosen appropriate to the particular test settings.
A requirement for applying Algorithm 3.2 is that the columns of the arising matrices $U_1$ and $U_2$ are linearly independent. Although a check of this necessity is not implemented directly, we note that it is hardly probable that linear dependency happens for very low-rank matrices. This expectation was confirmed by early test runs, where only rarely warnings were displayed in case of using the L-BFGS-update with a memory of at least 10.
Since the limited memory BFGS-updates are only well-defined if $(s^k)^T y^k > 0$ (where $s^k = x^{k+1} - x^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ holds), it is common to skip the update of the limited memory matrices if

$$(s^k)^T y^k < \varepsilon \|s^k\|^2. \tag{6.1}$$

For the SR1-update ill-conditioned steps are skipped automatically as described in Section 3.3.2 if we replace the definitions of the sets $\mathcal{I}_+$ and $\mathcal{I}_-$ by $\mathcal{I}_+ := \{i : \lambda_i > \varepsilon\}$ and $\mathcal{I}_- := \{i : \lambda_i < -\varepsilon\}$ for some tolerance $\varepsilon > 0$, see also the discussion in [39]. We choose $\varepsilon = 10^{-8}$ in the following experiments. The initial estimate $\gamma_k$ for the computation of the limited memory quasi-Newton matrices is set to

$$\gamma_k = \frac{(y^k)^T y^k}{(s^k)^T y^k},$$

which is suggested by Liu and Nocedal in [103] and links our methods to methods making use of Barzilai-Borwein techniques [11]. In particular, using a memory of zero, we arrive at first-order proximal methods and note that RPQN with a memory of zero has similarities with SpaRSA [164].
It remains to mention that there are some variants for the updates of the limited memory quasi-Newton matrices $H_k$. In case of unsuccessful steps in Algorithm 5.1 or in proximal gradient steps in Algorithm 4.1, one well-known approach is to delete all previous information and restart with $H_k$ as a multiple of the identity matrix. Since the procedure described above worked quite well in most of our experiments, we do not discuss other variants.

**Termination Criterion** In view of the differences in the structure of the tested algorithms mentioned in Section 6.1.1, it is not expedient to test the algorithms with their original termination criteria. Therefore, to obtain more comparable results, an initial run with very high accuracy is performed for each test example to compute an approximation $\psi^*$ to the optimal function value. We note that this exists for the subsequent test examples. We terminate each of the algorithms when the value $\psi(x^k)$ in the current iterate $x^k$ satisfies

$$\frac{\psi(x^k) - \psi^*}{\max\{1, |\psi^*|\}} \leq \texttt{tol} \tag{6.2}$$

for a predefined tolerance $\texttt{tol} > 0$. The term on the left hand side is referred to as *objective value error* in the following.

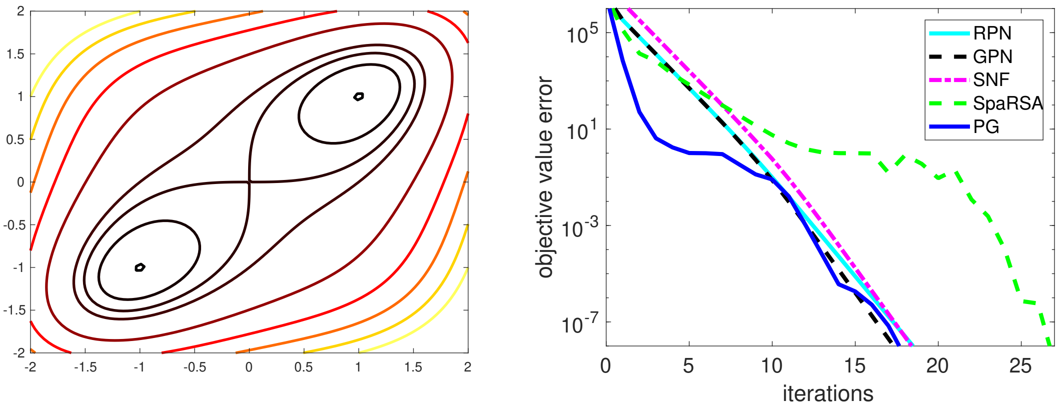
**Choice of Parameters** It remains to select the parameters for our methods. In GPN and QGPN, we use the parameters $p = 2.1$ and $\rho = 10^{-8}$ for the acceptance criterion (4.7). The line search is performed with $\beta = 0.1$ and $\sigma = 10^{-4}$. The constant $\tau_k$ for the proximal gradient step is initialized with $\tau_0 = 1/6$, and in each step adapted to reach the (local) Lipschitz constant of the gradient of $f$, similar to the technique described for PG in Section 6.1.1. The minimal and maximal value for this are $\tau_{\min} = 10^{-4}$ and $\tau_{\max} = 10^4$.
For RPN and RPQN we initialize the regularization parameter with $\mu_0 = 1$. The parameters for the acceptance criterion are set to $p_{\min} = 10^{-4}$ and $\kappa = 1.1$, and for the trust-region-type update we use $c_1 = 10^{-4}$, $c_2 = 0.9$, $\sigma_1 = 0.5$ and $\sigma_2 = 4$. The constants needed for the line search in RPQNm coincide with the above mentioned ones for QGPN.

## 6.2 A Simple Example

We start our analysis with the simple two-dimensional example given by

$$\min_{x,y} x^4 + y^4 - 4xy + 10^{-13}(|x| + |y|) \tag{6.3}$$

from [118]. This nonconvex objective function has three stationary points: a saddle point in $(0, 0)$ and two global minimizers with approximate coordinates $(1, 1)$ and $(-1, -1)$. The level sets of the function are illustrated in Figure 6.1(a). Note that the objective function is strongly convex near the global minimizers and, hence, the preliminaries of our theoretical analysis in Chapters 4 and 5 are fulfilled. Furthermore, the Hessian of the smooth part is easy to compute, so we do not apply any quasi-Newton strategy for this toy example. Instead, our focus is on the correlation between the number of iterations and the objective value error defined in (6.2). For that purpose, we disregard the effort of solving subproblems for the moment, which obviously differs for each of the considered methods.
We start with $x^0 = (30, 40)$, terminate if (6.2) holds with $\texttt{tol}= 10^{-8}$ and consider our second-order proximal methods GPN and RPN, the first-order proximal methods PG and SpaRSA and the semismooth Newton method SNF, which are described in Section 6.1. The modified method RPNm is not listed, since it turns out to coincide with RPN here. First, note that all methods finally reach one of the global minimizers, whereas none of them breaks down at the saddle point $(0, 0)$. While SpaRSA finds the minimizer near $(-1, -1)$, it is not surprising that most methods end up in the minimizer near $(1, 1)$ due to the choice of the starting point $x^0$.
The numerical performance of the methods is displayed in Figure 6.1(b). Noting again, that temporarily we leave out the costs of the individual iterations in the interpretation, there is



(a) Illustration of the level sets of the objective function in (6.3)

(b) Comparison of different methods for solving the problem

Figure 6.1: Two dimensional nonconvex example from Section 6.2

almost no difference in the performance of all second order methods. In contrast, SpaRSA needs significantly more iterations to hit the chosen tolerance. It is noteworthy that the proximal gradient method eventually performs as well as the tested second order methods, while it is even better for large tolerances. However, the latter may be due to the specific example. Taking the costs for evaluating the proximity operators in the subproblems into account, this very likely shows that the overall performance of PG is the best one applied to problem (6.3).

Altogether, this example illustrates that the efficient solution of the subproblems (1.2) is particularly important, which is generally known to be much more expensive for second-order methods than for first-order methods.

## 6.3 Logistic Regression with $\ell_1$-Penalty

In this example, which originates from [82], we consider the logistic regression problem

$$\min_{y,v} \frac{1}{m}\sum_{i=1}^{m} \log\Big(1+\exp\big(-b_i(a_i^T y+v)\big)\Big)+\lambda\|y\|_1, \tag{6.4}$$

where $a_i \in \mathbb{R}^n$ $(i=1,\dots,m)$ are given feature vectors, $b_i \in \{\pm 1\}$ are the corresponding labels, $\lambda > 0$, $y \in \mathbb{R}^n$, and $v \in \mathbb{R}$. Usually, we have $m \gg n$. Logistic regression is typically used to separate given data by a hyperplane as described in Example 1.2, see [88] for further information.

With $\phi\colon \mathbb{R} \to \mathbb{R}$, $\phi(u) := \log\big(1+\exp(-u)\big)$, $x := (y^T, v)^T$ and $A \in \mathbb{R}^{m\times(n+1)}$, where the $i$-th row of $A$ is $(b_i a_i^T, b_i)$ for $i=1,\dots,m$, we can write (6.4) equivalently as

$$\min_x \psi(x) := \frac{1}{m}\sum_{i=1}^{m} \phi\big((Ax)_i\big)+\lambda\|x_{[\{1,\dots,n\}]}\|_1. \tag{6.5}$$

The function $\phi$ is convex and strongly convex on any compact, convex set, and its derivative is globally Lipschitz continuous. Thus, this also holds for the smooth part of $\psi$ assuming that $A$ has full rank. Following the discussion in Section 5.3, Assumption 5.11(a) is fulfilled here and the convexity of the objective function yields Assumption 5.11(b). Thus, the objective function in (6.4) also has the KL-property and the local convergence theories of Chapters 4 and 5 apply (of course depending on the choice of the matrices $H_k$ to satisfy the Dennis-Moré-condition).

Although the Hessian of $f$ can be computed analytically, the computation costs for matrix-vector-multiplications with this matrix are quite high. Taking this into account, the analysis in this section focusses on approximating $H_k$ using limited memory quasi-Newton matrices. After giving the details of the tested examples, we discuss several issues regarding the test runs. These include an investigation of the inexactness condition (4.6) of Algorithm 4.1, a comparison of solvers for the subproblems in our algorithms, and a discussion of the performance of several methods compared to our QGPN and RPQN.

### 6.3.1 Algorithmic and Numerical Details

As described in Section 6.1.2, the crucial part of the implementation is the soution of the subproblems (1.2). This circumstance is discussed in detail in the following. Depending on the limited memory quasi-Newton update, we label the corresponding methods with the abbreviations L-BFGS or L-SR1 when using Algorithm 3.2. Furthermore we set the memory to 10. Since this method is not compatible with the inexactness condition (4.6) without some extra computations, we also investigate solving the subproblems with FISTA and SNF. In both cases, the initial point for the subproblem solvers is the current iterate $x^k$.

The performance of our second order methods QGPN and RPQN is discussed in comparison with the methods SNF, FISTA, SpaRSA and PG described in Section 6.1.1.

We follow the example in [33] and generate test problems with $n = 10^4$ features and $m = 10^6$ training sets. Each feature vector $a_i$ has approximately 10 nonzero entries, which are generated independently from a standard normal distribution. We choose $y^{\text{true}} \in \mathbb{R}^n$ with 100 nonzero entries and $v^{\text{true}} \in \mathbb{R}$, which are independently generated from the standard normal distribution and define the labels as

$$b_i = \operatorname{sign}\left(a_i^T y^{\text{true}} + v^{\text{true}} + v_i\right),$$

where $v_i \in \mathbb{R}$ $(i = 1, \dots, m)$ are chosen independently from a normal distribution with variance 0.1. The regularization parameter $\lambda$ is set to $0.1\lambda_{\max}$, where

$$\lambda_{\max} = \frac{1}{m}\left\| \frac{m_-}{m} \sum_{i:b_i=1} a_i + \frac{m_+}{m} \sum_{i:b_i=-1} a_i \right\|$$

is the smallest value such that $y^* = (0, v^*)$ is a solution of (6.4). Thereby, $m_+$ and $m_-$ denote the number of indices such that $b_i = +1$ or $b_i = -1$, respectively. The derivation of this value can be found in [88]. For all methods, we start with the initial value $x^0 = 0$.

We terminate each of the tested methods as soon as the objective value error in the current iterate $x^k$ satisfies (6.2) with `tol` $= 10^{-6}$ or the relative distance of consecutive iterates, i.e. the term

$$\frac{\|x^{k+1} - x^k\|}{\|x^k\|}, \tag{6.6}$$

falls below $10^{-12}$. Furthermore, we perform a maximum of 100 iterations for the second-order methods and a maximum of 1000 iterations for the first-order methods.

| method | term.-<br>crit. | iter | Newton-<br>iter | sub-<br>iter | function<br>eval | proximity<br>eval | matrix-vector<br>products |
|---|---|---|---|---|---|---|---|
| QGPN-F | max. | 29.1 | 29.1 | 2 015 | 30.2 | 2 471 | 58.3 |
| | dim. | 27.5 | 27.5 | 1 522 | 28.6 | 3 369 | 55.1 |
| | const. | 28.8 | 28.8 | 2 778 | 30.0 | 6 234 | 57.8 |
| QGPN-S | max. | 21.6 | 21.6 | 36.2 | 22.7 | 58.9 | 43.3 |
| | dim. | 24.6 | 24.6 | 115 | 25.7 | 279 | 49.2 |
| | const. | 28.6 | 28.6 | 1684 | 29.8 | 4049 | 57.4 |

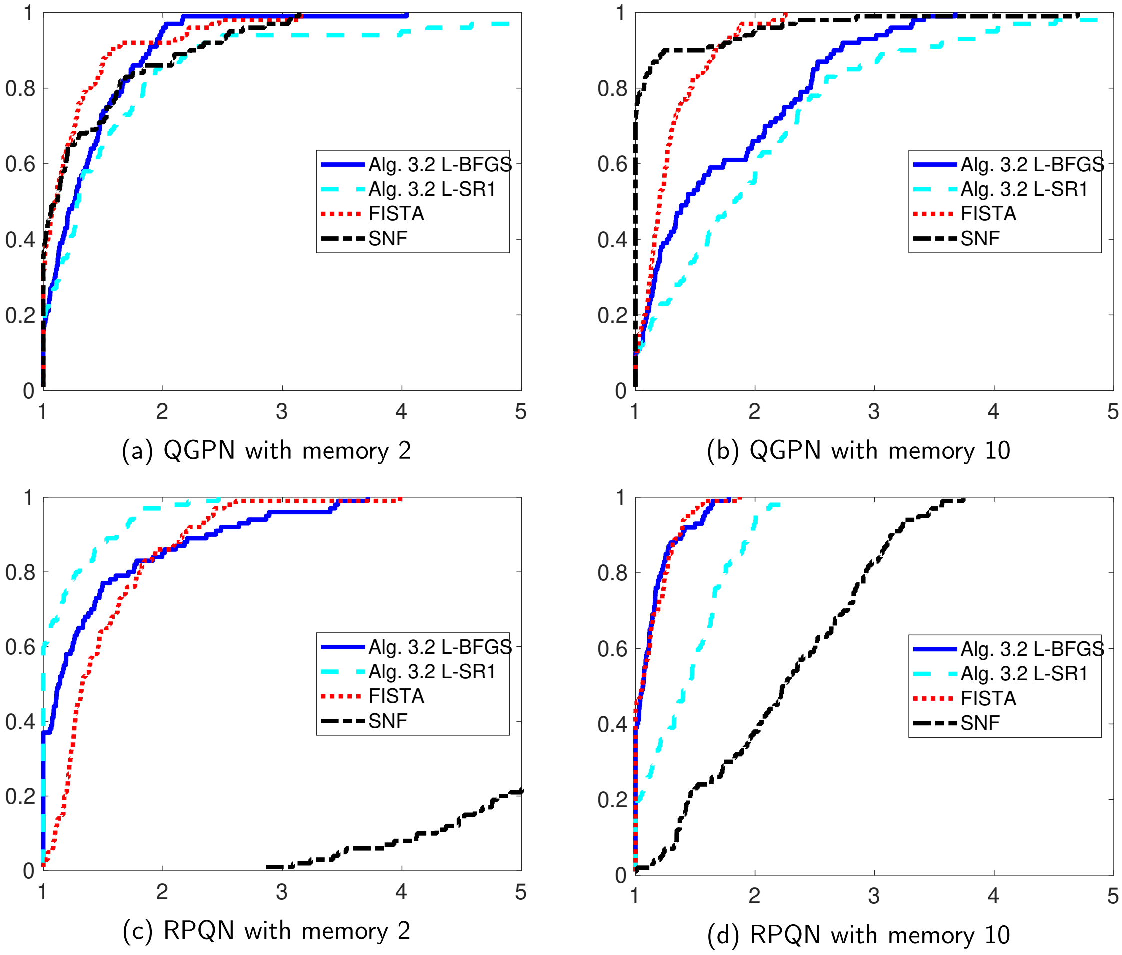
Table 6.1: Averaged values of 100 runs for the example in Section 6.3.2 with tolerance $10^{-6}$. Abbreviations: term.-crit. (method to terminate the solver for subproblems), iter (total number of (outer) iterations), Newton-iter (number of Newton-iterations), sub-iter (number of inner iterations), function eval (number of evaluations of the function $f$ or its gradient), proximity eval (number of evaluations of the proximity operator), matrix-vector products (number of evaluations of products $A \cdot x$ or $A^T \cdot x$) .

### 6.3.2 Inexactness Criteria for Solving Subproblems

We start with an investigation of the termination of the subproblems (1.2) with respect to the inexactness criterion (4.6) in Algorithm 4.1. As a consequence of Theorem 4.13, we can choose the sequence $\{\eta_k\}$ to be constant (const.). For our experiments, we computed an upper bound for $\overline{\eta}$ taking the constants in Theorem 4.13 and the analysis leading to this result, and set $\eta_k = 0.9\overline{\eta}$. A second possibility is to use a diminishing (dim.) sequence $\{\eta_k\}$. Here we investigated the sequence $\eta_k = 1/(k+1)$. Since the inexact termination criterion (4.6) is not practicable without significant additional computation costs (in particular for the evaluation of the proximity operator), we also test a third variant (max.): We minimize (1.2) using the standard termination criterion for the used solvers with a low maximal number of iterations, more precisely, 80 iterations for FISTA and 10 iterations for SNF, which resulted in the best performance in our experiments. The tolerance is adapted in each step such that the subproblems are solved more exactly when the current iterate is near the solution.

The averaged results of 100 runs for the described variants of our method are listed in Table 6.1. Looking at the variant with subproblem solver SNF, the computation costs using the diminishing or constant sequence $\{\eta_k\}$ are much higher than the costs using a maximum of 10 iterations. This can be seen looking at the outer iterations, but in particular in view of the numbers of inner iterations and evaluations of the proximity operator. Especially the number of evaluations of the proximity operator illustrates the difference in computation costs using the inexactness criterion in (4.6) and the approximation of the criterion by limiting the number of inner iterations. This is reasonable since there is one extra computation of the proximity operator in every inner iteration to check the inexactness condition. In contrast, the numbers of iterations are within the same range. Using FISTA to solve the subproblems, we observe a similar behaviour, although it is less marked here. We note that using GPN instead of QGPN leads to similar observations, looking at the major computation costs in the number of proximity evaluations, but the computational effort is significantly higher due to the number of matrix-vector-products involving the matrix $A$ in (6.5), cf. the corresponding numerical results in [82].

To draw a conclusion from these observations, when using the subproblem solvers FISTA or SNF in the following examples, we restrict the experiments and only investigate solving the subproblems with a maximum of 10 iterations (SNF) and 80 iterations (FISTA), as this choice leads to the lowest computation costs.



(a) QGPN with memory 2

(b) QGPN with memory 10

(c) RPQN with memory 2

(d) RPQN with memory 10

Figure 6.2: Performance profiles showing the runtime for 100 random test examples comparing different subproblem solvers as described in Section 6.3

### 6.3.3 Comparison of Subproblem Solvers

The aim of this section is the investigation of the performance of our methods QGPN and RPQN with various solvers for the solution of the subproblems (1.2). For both methods, we test FISTA and SNF as described in the previous section with a maximum of 80 and 10 iterations, respectively. In both cases, the quasi-Newton matrix $H_k$ is updated using the limited memory BFGS approach. In addition, the subproblems are solved using Algorithm 3.2 with both, the limited memory BFGS- and SR1-updates.

To accomplish comparability of these tests, we look at the runtime of the methods with the tested subproblem solvers considering 100 random test examples and document the results using the performance profiles introduced by Dolan and Moré [57] in Figure 6.2.

We note that in almost every tested example the numbers of outer iterations are almost identical, so the differences in the runtime indeed originate from the performance of the subproblem solvers. By considering the subproblem solvers in QGPN in Figure 6.2(a) and (b), we see that SNF and FISTA yield the best results, but the differences are small for a memory size of 2. On the other hand, for the larger memory size Algorithm 3.2 is outperformed. The reason is probably that the dimension of the semismooth system of equations in Algorithm 3.2, and thus the computational cost of solving it, increases with the size of the memory. In contrast, although the cost of matrix-vector multiplications

| method | iter | Newton-iter | succ. iter | sub-iter | function eval | proximity eval | matrix-vector products |
|---|---|---|---|---|---|---|---|
| RPQN (L-BFGS) | 43 | - | 35 | 42 | 44 | 121 | 88 |
| RPQN (L-SR1) | 20 | - | 20 | 17 | 21 | 56 | 42 |
| QGPN (L-BFGS) | 26 | 25 | - | 25 | 27 | 52 | 54 |
| QGPN (L-SR1) | 23 | 23 | - | 27 | 30 | 55 | 54 |
| SNF | 20 | 17 | - | 147 | 48 | 43 | 373 |
| FISTA | 463 | - | - | 531 | 1458 | 531 | 2453 |
| SpaRSA | 127 | - | - | 327 | 328 | 327 | 656 |
| PG | DNC | - | - | - | DNC | DNC | DNC |

Table 6.2: Values for the example in Section 6.3 with tolerance $10^{-6}$. In addition to the abbreviations from Figure 6.1 succ. iter is the number of successful or highly successful iterations performed by RPQN. DNC stands for 'does not converge' within the maximum iteration number.

also increases in SNF and FISTA, the dimension of the problem to be solved does not change with larger memory, which in turn does not increase the cost of solving it that significantly. Although a similar effect is observed for RPQN in Figure 6.2(c) and (d), SNF is clearly outperformed here and the performance of Algorithm 3.2 is really good. This might be caused by the different structure of the subproblems in contrast to QGPN, where no regularization term is used.

In summary, we see that for the considered class of examples the reduction of the subproblem (1.2) to a small-dimensional semismooth Newton system, using Algorithm 3.2 for solving the subproblems, yields a benefit in computation time in contrast to FISTA and SNF, especially for RPQN. For both methods, this advantage is bigger for small memory, while it reduces with larger memory. The experiments in the following section use RPQN and QGPN with Algorithm 3.2 to solve the subproblems. Although this is not visible directly in Figure 6.2, the best overall performance in the tests was achieved with a memory of 10. Hence, this is implemented for the subsequent examples.

### 6.3.4 Performance of Various Methods

To get an impression of the performance of the different algorithms in the considered problem setting, we first look at the objective value errors in relation to the respective runtime for one fixed example. Thereby, the time is averaged over 10 program runs to avoid the impact of first-time computation costs. Results are shown in Figure 6.3(a) and detailed data is collected in Table 6.2.

Looking at the performance in Figure 6.3(a), it is obvious that the second-order methods perform significantly better than the first-order methods. Due to the chosen size of the example this is not surprising. Looking at the first-order methods, the performance of PG is not convincing, and it does not reach the requested accuracy within 1000 iterations. However, also the accelerated methods FISTA and SpaRSA are outperformed by our methods.

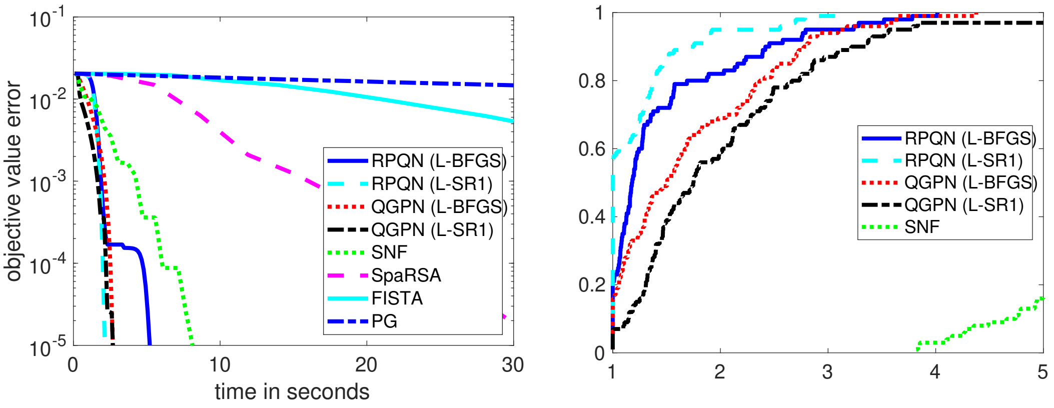
Note that the relation of the performance of SNF and QGPN is not comparable to the results in [82], since the algorithms have been improved. Furthermore, we mention that SNF is terminated, because the distance of two consecutive iterates is too small, whereas all other methods finally fulfill (6.2). The performance of our proximal quasi-Newton methods exceeds the one of SNF. The values in Table 6.2 imply that this is due to the costs of solving the subproblems. Furthermore, instead of using a quasi-Newton approximation, this method uses the exact Hessian. However, switching to an approximation might yield less matrix-vector-multiplications, but does not reduce the number of iterations for solving the subproblems. Hence, we did not consider that variant here.

Moreover, comparing the performance of RPQN and QGPN, there are only small differences except the outlier in RPQN (L-BFGS), where some intermediate unsuccessful iterations reduce the convergence speed. In addition, it is not unexpected that at the start QGPN is faster with reaching smaller errors, since RPQN might perform initial unsuccessful iterations to determine an appropriate value of $\mu_k$.

To accomplish better comparability of these methods, we look at the runtime of the tested second-order methods considering 100 random test examples and document the results using again the performance profiles from Dolan and Moré [57]. Results are shown in Figure 6.3(b).

Although SNF was originally developed for solving $\ell_1$-regularized problems [112], these thorough data show that SNF is clearly outperformed by our globalized proximal methods in the setting of this section. Moreover, we note that the performance of RPQN is better than the one of QGPN, but the differences are less significant. In particular, for RPQN the use of the limited memory SR1-method to update the matrices $H_k$ yields better results, whereas for QGPN the results with limited memory BFGS-matrices is superior.

Similar to the data presented in Table 6.2 the differences in the performance of the RPQN and QGPN methods result from the number of outer iterations and the subproblems are mostly solved within 1 or 2 steps. Taking the results of the previous section into account, the performance of QGPN might be slightly better when applying FISTA for solving the subproblems. Furthermore, almost all solutions of the subproblems of QGPN satisfy the descent condition (4.7), and, since the number of function evaluations is approximately equal to the number of outer iterations, almost all search directions are applied with full step length. Thus, for this example, the globalization by inserting proximal gradient steps is not necessary in practice. Since problem (6.5) is globally strongly convex if $A$ has full rank, a slight adaptation of our local convergence theory in Section 4.4 proves convergence even without this globalization. Details of this approach are left to the reader.



(a) Objective value error related to the averaged computation time for a random test example over 10 test runs

(b) Performance profile showing the runtime for 100 random test examples

Figure 6.3: Performance of the example described in Section 6.3 for logistic regression with $\ell_1$-penalty

## 6.4 Least Squares Problems with Group Sparse Regularizer

In this section, which is mainly taken from [81], we consider the least squares problem described in Example 1.1 for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ to find an approximate solution of the linear system $Ax = b$. To ensure sparsity of the solution, we use the $\ell_1$-$\ell_2$-sparsity regularizer, which is also called group sparse regularizer in the literature, cf. Example 1.2. The problem is given by

$$\min_x \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_{2,1} \tag{6.7}$$

for some $\lambda > 0$, where

$$\|x\|_{2,1} := \sum_{j=1}^{p} \|x_{[\mathcal{I}_j]}\|.$$

Here, the index sets $\mathcal{I}_j$ $(j = 1,\dots,p)$ form a partition of $\{1,\dots,n\}$. Since the sets $\mathcal{I}_j$ are pairwise disjoint, the proximity operator $\mathrm{prox}_{\lambda\|\cdot\|_{2,1}}$ and a Newton derivative of it can be computed block-wise using Proposition 3.1 as well as the formulas in Proposition 3.2 and Example 3.23. The use of the $\ell_1$-$\ell_2$-regularizer makes sense in many applications, where sparsity should be achieved with respect to some groups of variables, which was addressed in Example 1.2.

The gradient $\nabla f(x) = A^T(Ax-b)$ of the function $f(x) = \frac{1}{2}\|Ax-b\|^2$ is obviously Lipschitz continuous. Hence, the assumptions of the global convergence theorems for Algorithm 4.1 (Theorems 4.3 and 4.4) and Algorithm 5.1 (Theorem 5.7) are satisfied. Furthermore, the objective function $\psi$ is convex and coercive here, so that the set of solutions is nonempty and bounded. Thus, a sequence $\{x^k\}$ generated by one of these methods has at least one accumulation point, which is a global minimizer of the objective function, and the sequence $\{\psi(x^k)\}$ converges to the optimal function value. Similar to the discussion in the previous section we see that the preliminaries of the local convergence theories in Chapters 4 and 5 apply.

Taking the high costs of matrix-vector-multiplications with the matrix $A$ and Remark 2.50 into account, the analysis in this section again focusses on approximating $H_k$ using limited memory quasi-Newton matrices. The algorithmic details are almost the same as in the previous section and are therefore not repeated. We continue with the details of the implemented test setting, before investigating several topics regarding the test runs. These are a comparison of different sizes of memory for the quasi-Newton update, some notes on the effect of the modification of RPQN (Algorithm 5.2), and a discussion of the performance of several methods compared to our second-order algorithms.

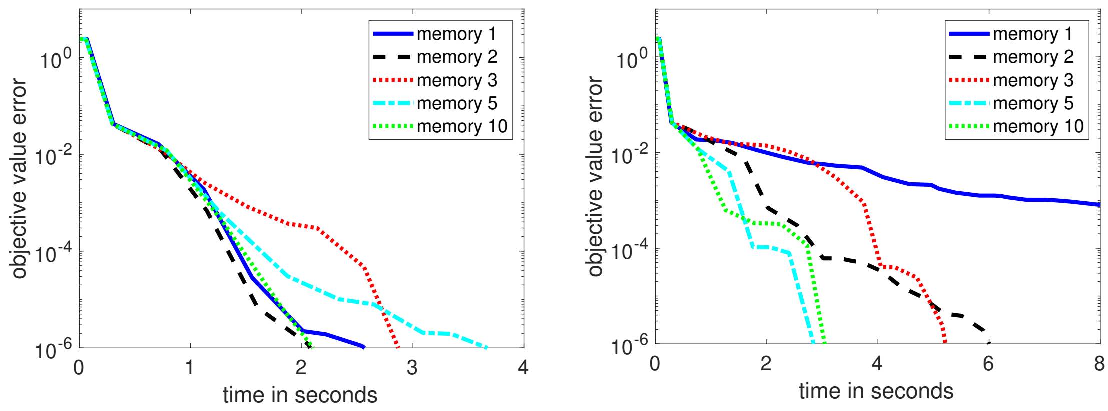
### 6.4.1 Problem Setting and Implementation

We follow the generic example in [16] and choose the entries in $A$ and $b$ from a uniform distribution in $[0, 1]$ with $n = 2500$ and $m = 1600$. (In Section 6.4.4 these are replaced by $n = 25k$ and $m = 16k$ for various values $k \in \mathbb{N}$.) The parameter $\lambda$ is set to 1. Furthermore, the index sets $\mathcal{I}_j$ are chosen randomly with 4 to 12 elements. We start with the initial guess $x^0 = 0$.

The focus of this section is the investigation of the performance of RPQN. For that purpose, we look at RPQN and RPQNm, where the subproblems are solved using Algorithm 3.2 with limited memory BFGS- and SR1-updates. With regard to Remark 2.50, QGPN plays a minor role here. For this method, subproblems are also solved using Algorithm 3.2 with limited memory BFGS-updates and a memory of 5, which yielded the best performance in test runs. Besides these methods, we compare the performance to SNF, FISTA, SpaRSA

| method (memory) | iter | highly s. iter | succ. iter | unsucc. iter | sub-iter | function eval | proximity eval | matrix-vector products |
|---|---|---|---|---|---|---|---|---|
| L-BFGS (1) | 46 | 18 | 14 | 14 | 199 | 47 | 442 | 94 |
| L-BFGS (2) | 36 | 18 | 5 | 13 | 149 | 36 | 333 | 73 |
| L-BFGS (3) | 49 | 27 | 6 | 16 | 208 | 50 | 461 | 100 |
| L-BFGS (5) | 55 | 32 | 3 | 20 | 265 | 53 | 577 | 106 |
| L-BFGS (10) | 34 | 20 | 2 | 12 | 121 | 33 | 276 | 66 |
| L-SR1 (1) | 568 | 339 | 54 | 175 | 2692 | 500 | 5789 | 1000 |
| L-SR1 (2) | 92 | 57 | 3 | 32 | 464 | 87 | 1000 | 174 |
| L-SR1 (3) | 76 | 47 | 3 | 26 | 359 | 75 | 780 | 150 |
| L-SR1 (5) | 45 | 27 | 2 | 16 | 206 | 45 | 453 | 90 |
| L-SR1 (10) | 49 | 30 | 2 | 17 | 207 | 48 | 458 | 96 |

Table 6.3: Values for the example in Section 6.4 using RPQN with different memories and tolerance $10^{-6}$. In addition to the abbreviations from Figure 6.1 highly s. iter, succ. iter, and unsucc. iter are the numbers of highly successful, successful and unsuccessful iterations performed by RPQN, respectively.



(a) RPQN with limited memory BFGS

(b) RPQN with limited memory SR1

Figure 6.4: Error plot for RPQN with limited memory quasi-Newton approach and different memories for the setting in Section 6.4.

and PG as described in Section 6.1.1.

As before, we terminate the methods if the current iterate satisfies (6.2) with tol$= 10^{-6}$, or the relative distance (6.6) of two consecutive iterates is less than $10^{-10}$, where we obviously consider consecutive successful or highly successful iterates for RPQN.

### 6.4.2 Effect of the Memory Size for the Limited Memory Quasi-Newton Approach

An important aspect in the study of limited memory quasi-Newton methods is the size of the memory. For investigating this purpose, we consider the regularized proximal quasi-Newton method RPQN and update the matrix $H_k$ by the limited memory BFGS- and the limited memory SR1-approach explained in Section 3.3.1. We use one example in the above test setting and run the methods with memory sizes of mem$= 1, 2, 3, 5, 10$. Results are reported in Table 6.3. In Figure 6.4 we display the objective value error defined in (6.2) relative to the computation time. To avoid the effect of first-time computation costs, we use the mean of 10 test runs and note the different scaling of the time axis in the Figures 6.4(a) and (b). We first investigate the performance of using RPQN with the limited memory BFGS-approach. Figure 6.4(a) shows that the overall performance is almost identical for a memory of 1,2 and 10, but the difference to the memories 3 and 5 is not large. For a more detailed analysis consider the corresponding rows in Table 6.3. While the values for `mem`= 2 and `mem`= 10 almost coincide, the test run with `mem`= 1 performs significantly more successful (not highly successful) iterations and the listed values are higher than for the previously mentioned memory sizes. Thus, the performance for `mem`= 1 in Figure 6.4(a) must result from lower computation costs for solving the subproblems. The behaviour for `mem`= 3 and `mem`= 5 originates from the higher number of (outer) iterations and, thus, also a higher number of inner iterations for solving the subproblems.

For the limited memory SR1-approach, results are different. In particular, the method with a memory of 1 is not satisfactory, which presumably is a consequence of poor results in the subproblems. On the other hand, the performance for `mem`= 5 and `mem`= 10 is almost the same, both, with view to Figure 6.4(b) and Table 6.3. Hence, the positve effect of choosing a larger memory has a stronger impact than the additional costs for solving the subproblems, at least until a memory size of 5.

Overall, it is remarkable that the algorithms perform almost exclusively highly successful or unsuccessful iterations, while hardly any are just successful. Following the results of this section, we use `mem`= 10 for the subsequent tests.

### 6.4.3 Numerical Impact of the Modified RPQN

In Section 5.5 we introduced a modified version of the regularized proximal Newton-type method. While the original algorithm (Algorithm 5.1) simply continues with the next iteration in case of unsuccessful steps, its modified version (Algorithm 5.2) applies a proximal gradient approach in that case. This means that whenever an iterate is unsuccessful, instead a proximal gradient step combined with an Armijo-type line search is performed similar to Algorithm 4.1. As deduced in Section 5.5, this change does not yield any theoretical benefits. Hence, the purpose of this section is to investigate the numerical advantages.

Similar to the previous section we run RPQN and RPQNm updating the quasi-Newton matrix with the L-BFGS- and L-SR1-approach. The objective value error (6.2) relative to the average computation time of 10 test runs of a random example as described at the beginning of Section 6.4 is displayed in Figure 6.5 and detailed data of the experiment are given in Table 6.4.

For both, the limited memory BFGS- and the limited memory SR1-approach, we see a significant improvement when using RPQNm. In detail, the number of iterations considerably reduces. For the tested example the reason is easy to see: The modified algorithm performs one unsuccessful iteration and therefore one proximal gradient step, which brings the iterate to a point, from which all remaining iterations are highly successful. Therefore, also all values documented in Table 6.4 are significantly lower than for RPQN. It is remarkable

| method | iter | highly<br>s. iter | succ.<br>iter | unsucc.<br>iter | sub-<br>iter | function<br>eval | proximity<br>eval | matrix-vector<br>products |
|---|---|---|---|---|---|---|---|---|
| RPQN | 34 | 20 | 2 | 12 | 121 | 33 | 276 | 66 |
| RPQNm | 11 | 10 | 0 | 1 | 70 | 19 | 154 | 38 |
| RPQN | 49 | 30 | 2 | 17 | 207 | 48 | 458 | 96 |
| RPQNm | 9 | 8 | 0 | 1 | 50 | 17 | 112 | 34 |

Table 6.4: Values for the example in Section 6.4.3 comparing RPQN and RPQNm. The first two rows use L-BFGS, the others use L-SR1. The columns have the same meaning as in Table 6.3, see also Table 6.1.

that this behaviour is even more marked when the quasi-Newton matrices are updated by the limited memory SR1-approach. Overall, this section shows that the modification of RPQN in Algorithm 5.2 results in a substantially improved numerical performance.

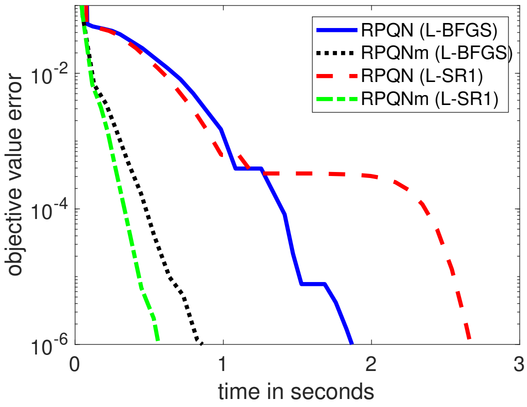


Figure 6.5: Convergence plot for the comparison of RPQN and RPQNm as described in Section 6.4.3.

### 6.4.4 Performance of Various Methods

After the detailed study of the two regularized proximal Newton-type methods from Chapter 5, we now turn to the comparison of these methods with QGPN as well as the state-of-the-art methods described in Section 6.1.1. For this purpose, we continue to consider problem (6.7) with the test setting described in Section 6.4.1. To get an accurate impression of the performance of the tested methods, we use $n = 25k$ for $k \in \{1, 3, 10, 30, 100, 300\}$ and a constant column-to-row ratio $m/n = 16/25$ and document in each case the average computation time of 10 test examples. The results are presented in Figure 6.6. For reasons of clarity we displayed the methods RPQN, RPQNm and QGPN only in combination with the limited memory SR1-approach, but note that the performance for the BFGS-approach is comparable.

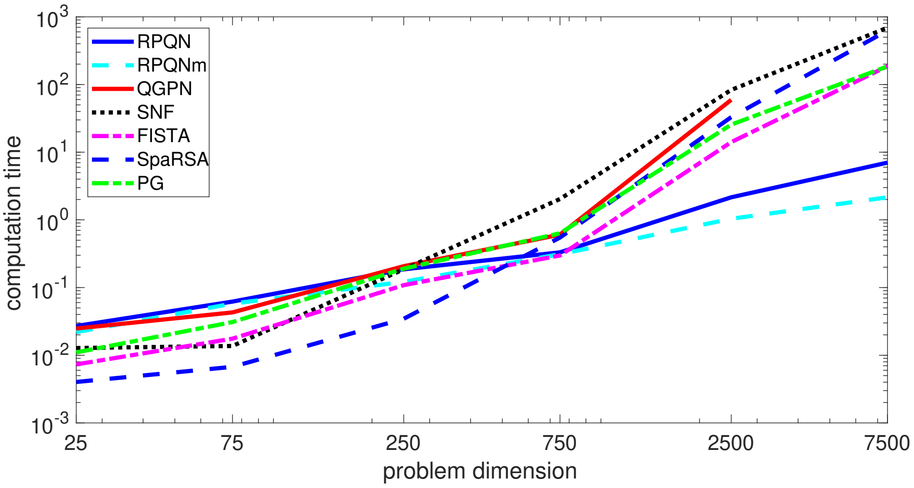


Figure 6.6: Comparison of the performance of several methods depending on the problem dimension as described in Section 6.4.4.

For small dimensions, we see at the left edge of Figure 6.6 that the first-order methods, led by SpaRSA, perform best. Of course, this is not surprising, since for small dimensions the cost of solving the subproblems is higher than its benefit and multiples of the identity matrix already provide workable approximations for the matrices $H_k$. That advantage disappears at a dimension of 750, where the tide turns and for bigger problem dimensions RPQN and RPQNm perform significantly better than the other methods. Given that this does not hold for the second-order methods SNF and QGPN, however, this observation is even more remarkable. In particular, QGPN does not even achieve the desired accuracy within the maximum iteration number for $n = 7500$.

A closer look to the investigated first-order methods also shows that SpaRSA is particularly useful for smaller dimensions in this test setting, while FISTA performs better for larger ones. While PG is mostly slightly worse than FISTA, both methods show similar behavior for $n = 7500$. Presumably, further investigation of the performance of PG would expectedly show that PG will keep up or even overtake the performance of FISTA for even larger dimensions.

## 6.5 Student's $t$-Regression with $\ell_1$-Penalty

The aim of many applications of inverse problems is to find a preferably sparse solution $x^* \in \mathbb{R}^n$ of the problem $Ax = b$ with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, cf. Example 1.1. A general solution to this issue is to consider the Lasso-problem (least absolute shrinkage and selection operator)

$$\min_x \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1$$

for some $\lambda > 0$. If $b$ is not known exactly but only a noisy approximation $\hat{b} \approx b$, this works well if the error in the entries of $\hat{b}$ is Gaussian. Particularly, the impact of large errors is very large. The reduction of the influence of large errors motivates to replace the quadratic loss by the Student loss, see Figure 6.7(a), which yields the problem

$$\min_x \sum_{i=1}^m \phi\big((Ax-b)_i\big) + \lambda\|x\|_1 = \sum_{i=1}^m \log\left(1 + \frac{(Ax-b)_i^2}{\nu}\right) + \lambda\|x\|_1 \tag{6.8}$$

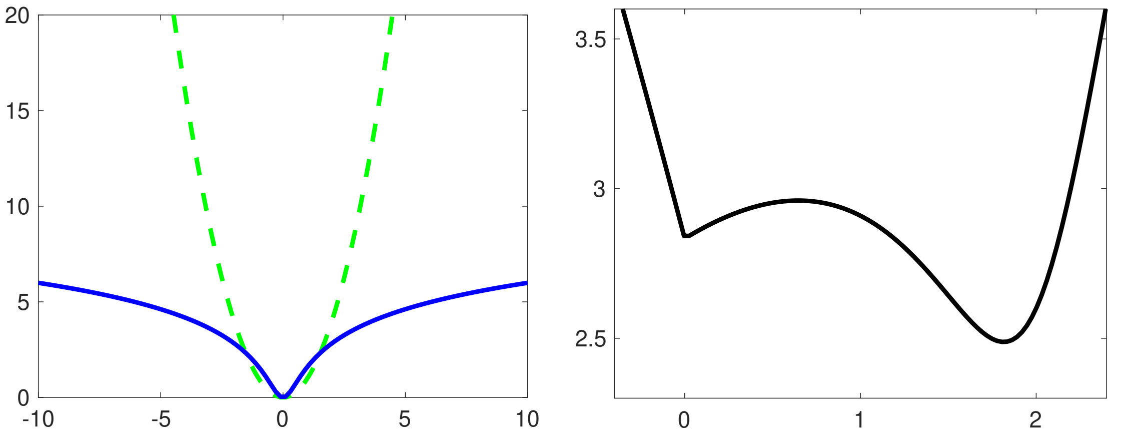
with $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(u) := \log(1 + u^2/\nu)$ for some $\nu > 0$, which is sometimes referred to as *degree of freedom*. For more information on Student's $t$-distribution, we refer to [3, 4, 112] and the references therein.

It is easy to see that the derivative of $\phi$ is Lipschitz continuous and $\phi$ is coercive, but not convex. Thus, the smooth part of $\psi$ has a (globally) Lipschitz continuous gradient. However, the lack of convexity makes many state-of-the-art methods inapplicable. Besides first-order methods like FISTA, we mention classical proximal Newton-type methods without globalization, e.g. [95]. Moreover, a minimizer of (6.8) is expected to approximately solve the linear system $Ax = b$. Since $\phi$ is strongly convex in $B_{\sqrt{\nu}}(0)$, we expect that in a solution of (6.8) the local convergence theory of Chapters 4 and 5 is applicable.

Before introducing two different problem settings and comparing the performance of several methods, we note that this section is mainly based on Section 5.2 in [82]. However, test runs are performed with revised and improved implementations of the algorithms.

### 6.5.1 Test Setting with low-cost matrix $A$

We first investigate the test setting decribed in [112]. Let $n = 512^2$ and $m = n/8 = 32768$. The matrix $A \in \mathbb{R}^{m\times n}$ takes $m$ random cosine measurements, i.e for a random subset



(a) Student's $t$-error compared to the Gaussian error function

(b) The function $\log\left(1+(x-2)^2/0.25\right)+1.3|x|$ in the form (6.8) has more than one local minimum

Figure 6.7: Properties of the Student's $t$-function introduced in Section 6.5

$\mathcal{I} \subset \{1,\ldots,n\}$ with $m$ elements, define $Ax = (\mathrm{dct}(x))_{[\mathcal{I}]}$, where dct denotes the discrete cosine transform.

We generate a true sparse vector $x^{\text{true}} \in \mathbb{R}^n$ with $k = \lfloor n/40 \rfloor = 6553$ nonzero elements, whose indices are chosen randomly. The nonzero components are computed via $x_i^{\text{true}} = \eta_1(i)10^{\eta_2(i)}$ with a random sign $\eta_1(i) \in \{\pm 1\}$ and $\eta_2(i) \in [0,1]$ chosen independently from a uniform distribution. The approximate image $b \in \mathbb{R}^m$ is generated by adding Student's $t$-noise with degree of freedom 0.25 and rescaled by 0.1 to $Ax^{\text{true}}$. Furthermore, we set $\nu = 0.25$ and $\lambda = 0.1\lambda_{\max}$, where $\lambda_{\max}$ is the critical value, for which $x^0 = 0$ is a critical point of (6.8). Using Fermat's rule (Proposition 2.47) applied to the objective function in (6.8), a short calculation proves

$$\lambda_{\max} = 2\left\| \sum_{i=1}^{m} \frac{b_i}{\nu + b_i^2} a_i \right\|_\infty ,$$

where $a_i^T$ is the $i$-th row of the matrix $A$.

We start with the initial point $x^0 = A^T b$ and, again, terminate each of the algorithms, when the value $\psi(x^k)$ in the current iterate $x^k$ satisfies (6.2) with `tol`$= 10^{-6}$. It is important to mention that due to the lack of convexity, functions in the form of problem (6.8) might have several local minima (and therefore several stationary points) with different function values, see the example in Figure 6.7(b). However, test runs showed that all methods, if they were convergent, finally reached a point with the same function value. Hence, the termination criterion is still applicable in this nonconvex problem setting.

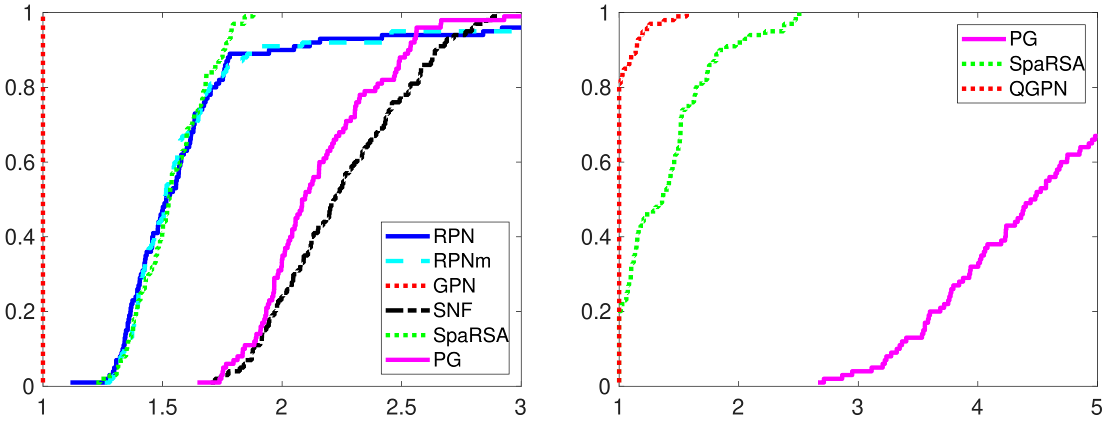
As the discrete cosine transform is a predefined Matlab-function, the computation costs for matrix-vector-products with $A$ or $A^T$ are significantly lower than computing products with limited memory quasi-Newton matrices. For this reason, we implemented the methods RPN, RPNm and GPN with $H_k = \nabla^2 f(x^k)$ instead of using appropriate approximations. The subproblems are solved using SNF, which turned out to be more efficient than using FISTA here.

In order to provide comparability, we look again at the runtime of 100 test examples and document the performance using the performance profile introduced in [57]. The results are shown in Figure 6.8(a) and numerical data are listed in Table 6.5. The first observation

| method | iter | Newton-<br>iter | succ.<br>iter | sub-<br>iter | function<br>eval | proximity<br>eval | matrix-vector<br>products |
|---|---|---|---|---|---|---|---|
| RPN | 55.1 | - | 54.9 | 108.2 | 55.9 | 228.3 | 1162.2 |
| RPNm | 56.1 | - | 56.0 | 110.1 | 57.1 | 233.7 | 1202.5 |
| GPN | 48.7 | 48.7 | - | 78.7 | 49.7 | 127.4 | 719.3 |
| SNF | 69.1 | 40.6 | - | 279.4 | 109.7 | 137.1 | 737.6 |
| SpaRSA | 553.3 | - | - | 553.3 | 554.3 | 553.3 | 1 108.6 |
| PG | 556.0 | - | - | - | 1112.9 | 556.0 | 1669.9 |

Table 6.5: Numerical data for the example in Section 6.5.1. The columns have the same meaning as in Table 6.2, see also Table 6.1.

is that GPN clearly outperforms all other methods in this setting. In detail, all iterations performed by GPN are Newton iterations and the number of function evaluations indicates that always the full step length is attained. Thus, the method performs very well in this nonconvex setting. Looking at RPN and RPNm, it can be seen that the modification is not useful for this example. Furthermore, in comparison to GPN, especially the high number of matrix-vector products including the matrix $A$ and almost twice as much evaluations of the proximity operator cause the behaviour of RPN, which shows a similar performance than the first-order method SpaRSA. Although the data listed for SNF in Table 6.5 is relatively small, this method is not convincing. Reasons might be the low number of Newton iterations and the higher costs of subiterations compared to the ones in our proximal Newton methods.



(a) Example from Section 6.5.1

(b) Example from Section 6.5.2

Figure 6.8: Performance profiles showing the runtime for 100 random test examples of Student's $t$-regression with $\ell_1$-penalty described in Section 6.5

### 6.5.2 Test Setting for the Limited Memory Quasi-Newton Approach

To demonstrate the performance of the limited memory quasi-Newton versions QGPN and RPQN of our algorithms, we construct a second test example with higher computation costs for the matrix-vector products with the matrices $A$ or $A^T$. For that purpose, we modify the setting of the previous section and use $A$ as defined in Section 6.3, this is $n = 10^4$, $m = 10^6$ and $A \in \mathbb{R}^{m \times n}$ having approximately 10 nonzero entries in every row. Moreover, we use the initial value $x^0 = 0$. Everything else remains unchanged.

For the comparison, we use the same methods as before, but utilize the quasi-Newton approaches QGPN instead of GPN and RPQN instead of RPN with a limited memory

BFGS-update for the Hessian-approximations $H_k$ and a memory of 10. The proximal Newton subproblems are solved using Algorithm 3.2. First, we note that the performance of RPQN and SNF is not satisfactory in the given setting and the convergence is very poor. While in SNF this might be due to the fact that this method is not a descent method, the performance of RPQN fails, since too many updates of the BFGS-matrix are skipped because (6.1) is violated. It might be a topic of future experiments to investigate in various update strategies of $H_k$ in this case to improve the convergence.

The performance of the remaining methods is shown in the performance profile in Figure 6.8(b), where the convergence properties of Figure 6.8(a) are confirmed. In detail, QGPN ist still by far the best method (note the different scaling of the two performance profiles in Figure 6.8) and the only second-order method with appropriate convergence results. Considering the first-order methods, PG is again outperformed by SpaRSA.

## 6.6 Nonconvex Image Restoration

So far the numerical results of this chapter were conducted with synthetic data. In this section, which is based on [81], however, we demonstrate the performance of our algorithms for a real-world problem of image restoration. Given a noisy blurred image $b \in \mathbb{R}^n$ and a blur operator $A \in \mathbb{R}^{n \times n}$, the aim is to find an approximation $x \in \mathbb{R}^n$ to the original image satisfying $Ax \approx b$. Note that, for simplicity in notation, we handle the images $x, b$ as vectors in $\mathbb{R}^n$.

Similar to the previous section, assuming that the noise in $b$ follows Student's $t$-distribution, cf. [4], this leads to the problem

$$\min_x \sum_{i=1}^n \phi\big((Ax-b)_i\big) + \lambda \|Bx\|_1 \tag{6.9}$$

with $\phi(u) := \log(1+u^2)$ and some $\lambda > 0$. In contrast to the previous sections, the nonsmooth term $\varphi(x) = \lambda\|Bx\|_1$ does not have the purpose to obtain sparsity of the solution, but to get smooth gradations and guarantee antialiasing in the final image. For that purpose, $B : \mathbb{R}^n \to \mathbb{R}^n$ is a two-dimensional discrete Haar wavelet transform. Haar wavelets were originally introduced in the analysis of signals, but have several more applications in the meantime [162].

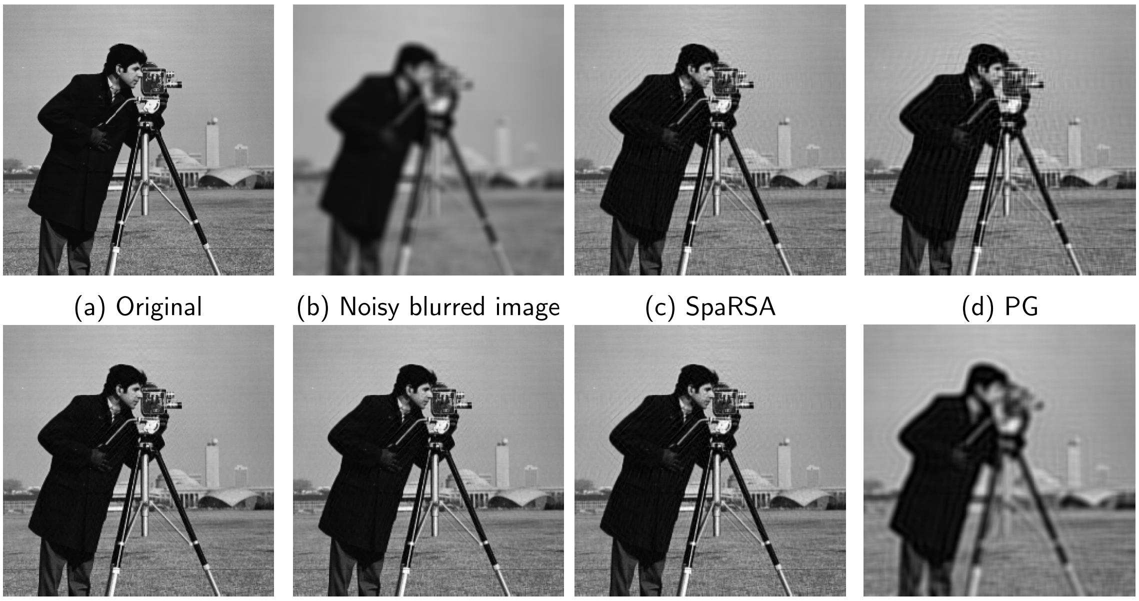
Since $B$ is orthogonal, the proximity operator of $\varphi$ is given by

$$\operatorname{prox}_\varphi(u) = B^T \operatorname{prox}_{\|\cdot\|_1}(Bu),$$

cf. Proposition 3.3, and, hence, can be computed analytically. This observation was used in [82] for numerical experiments. In this thesis, instead we make use of the orthogonality of $B$ to replace problem (6.9) with the equivalent formulation

$$\min_y \sum_{i=1}^n \phi\big((AB^Ty-b)_i\big) + \lambda\|y\|_1, \tag{6.10}$$

where $y = Bx$. Similar to the analysis in Section 6.5 we expect a solution $y^*$ of (6.10) to satisfy $AB^Ty \approx b$. Under this hypothesis $\phi$ is strongly convex in a neighbourhood of the minimizer and $\nabla\phi$ is Lipschitz continuous. Thus, the explanation in Section 5.3 yields that the requirements for the global and local convergence theories in Chapters 4 and 5 are satisfied.



(a) Original (b) Noisy blurred image (c) SpaRSA (d) PG

(e) RPQN (f) RPQNm (g) QGPN (h) SNF

Figure 6.9: Original and blurred image and recovered images using the stated algorithms as described in Section 6.6 (terminated after a computation time of 10 seconds)

We follow the test setting in [35], see also [82, 148], to restore a grayscale test image with $256 \times 256$ pixels, hence $n = 256^2 = 65536$. The mapping $A$ is a Gaussian blur operator of size $9 \times 9$ with standard deviation 4 and $B$ is the two-dimensional discrete Haar wavelet of level 4. Furthermore, we choose $\lambda = 10^{-4}$. The noisy blurred image $b$ is created from the original *cameraman* test image by applying $A$ and adding Student's $t$-noise with degree of freedom 1 and rescaled by $10^{-3}$. Finally, we initialize the tested methods with $y^0 = Bb$. For our analysis, we solve problem (6.10) using RPQN, RPQNm and QGPN with a limited memory SR1-update for $H_k$ and a memory of 2. For the sake of clarity we refrain from displaying results for a limited memory BFGS-update, but note that these results are comparable to the ones presented. Subproblems are again solved using Algorithm 3.2. In addition, we apply SNF, SpaRSA and PG to this test setting. Data of the tests are shown in Table 6.6, where the algorithms are terminated when (6.2) holds with `tol`$= 0$. In contrast to the previous experiments, $\psi^*$ is the value of the objective function (6.9) in the original image, which is not the optimal function value. Furthermore, we ran all algorithms 10 times to gain the average computation time of the methods. The objective value error (6.2) of these tests is displayed in Figure 6.11(a), where we note that SNF is not listed due to its poor performance. The resulting images after a runtime of 10 seconds are presented

| method | iter | Newton-iter | succ. iter | sub-iter | function eval | proximity eval | matrix-vector products |
|---|---|---|---|---|---|---|---|
| RPQN | 890 | - | 866 | 1790 | 891 | 4448 | 1790 |
| RPQNm | 883 | - | 869 | 1762 | 906 | 4423 | 1812 |
| QGPN | 1101 | 1098 | - | 1175 | 1113 | 2354 | 2215 |
| SNF | 183 | 91 | - | 1189 | 784 | 408 | 3855 |
| SpaRSA | 1089 | - | - | 1964 | 1965 | 1964 | 3930 |
| PG | 1269 | - | - | - | 2594 | 1269 | 3864 |

Table 6.6: Numerical data for the image restoration example in Section 6.6. The columns have the same meaning as in Table 6.2, see also Table 6.1.

in Figure 6.9, while enlarged image sections displaying a part of the tripod and the domed building in the background of the image are shown in Figure 6.10. Note that the results are minimal lighter than the original image. This is caused by the fact that we used the Haar wavelet of level 4 and not of the maximal level $\log_2(256) = 8$.

All of the presented results show that the performance of our second-order proximal quasi-Newton methods is outstanding compared to the other methods, see in particular the pictures in Figure 6.10. Although the test setting is identical to the one in Section 6.5.2, RPQN and RPQNm yield very good results and almost none of the updates of $H_k$ is skipped. Looking at the details of the reconstructed images in Figure 6.10, the performance of RPQNm is by far the best one, while also the results of RPQN and QGPN are satisfactory. Similar to the previous section we obtain that SNF does not yield an adequate performance. As described earlier, the problem with this method is probably that the semismooth Newton steps reduce $\|r(x^k)\|$ defined in (4.4), while probably increasing $\psi(x^k)$, whereas the proximal gradient steps used for the globalization decrease $\psi(x^k)$, but probably increase $\|r(x^k)\|$. For nonconvex problems, where these steps are expected to alternate steadily, this seems to result in a lack of performance.

On the other hand, the first order methods SpaRSA and PG yield adequate results of the reconstructed image, but the performance, especially with view to Figure 6.11(a), can not keep up with the tested second-order proximal methods.

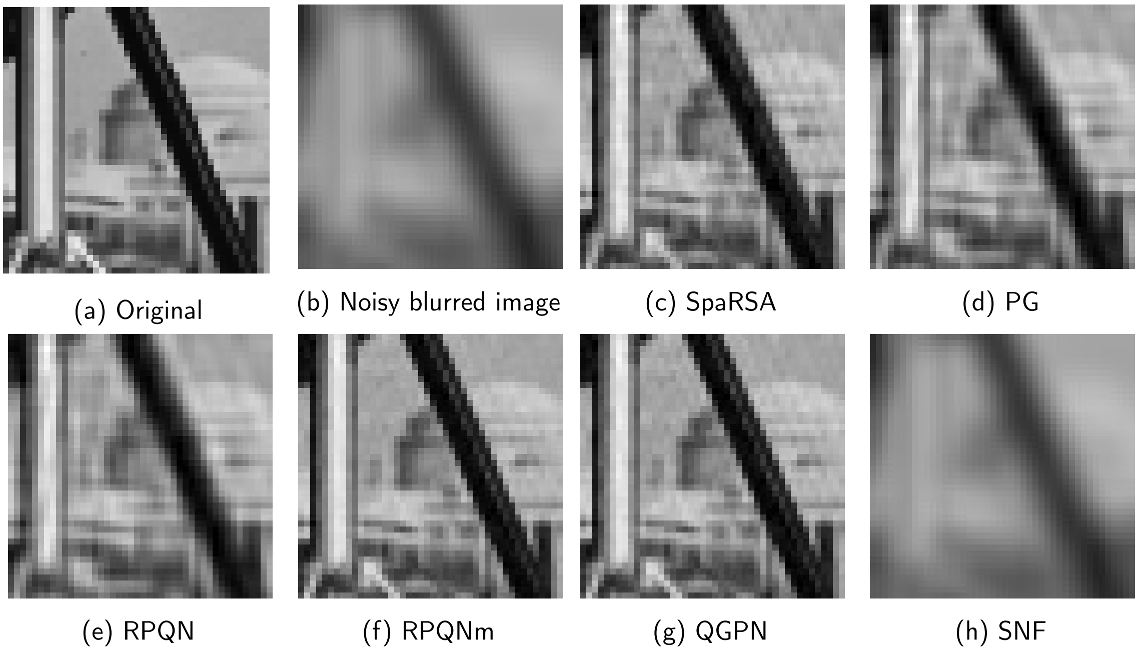

(a) Original (b) Noisy blurred image (c) SpaRSA (d) PG
(e) RPQN (f) RPQNm (g) QGPN (h) SNF

Figure 6.10: Original and blurred image and recovered images using the stated algorithms as described in Section 6.6 (terminated after a computation time of 10 seconds): Enlarged image detail from the results in Figure 6.9

## 6.7 Logistic Regression with Overlapping Group Regularizer

The focus of this thesis and the foregoing numerical examples lies on applications where the proximity operator $\text{prox}_\varphi^H$ of the nonsmooth function $\varphi$ in problem (1.1) can be computed analytically as long as $H$ is a positive multiple of the identity matrix. However, there are applications for which this is not the case. One of those is addressed in this section, which is based on the author's work in [82].

A main advantage of the globalized proximal Newton method (Algorithm 4.1) over semismooth Newton methods is that it is also applicable in the above described case that there is no known formula to compute the proximity operator of the nonsmooth function $\varphi$ and consequently no easy way to compute a Newton derivative thereof. A range of applications can be formulated in the form

$$\min f(x) + \tilde{\varphi}(Bx),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is still a continuously differentiable function, $\tilde{\varphi} : \mathbb{R}^m \to \overline{\mathbb{R}}$ is convex, proper and lower semicontinuous, and $B \in \mathbb{R}^{m\times n}$. Assuming that the proximity operator $\operatorname{prox}_{\tilde{\varphi}}$ and the matrix $B$ are explicitly available, several methods [30, 42, 43] are designed to exploit this structure.

As an example we consider the group penalty function

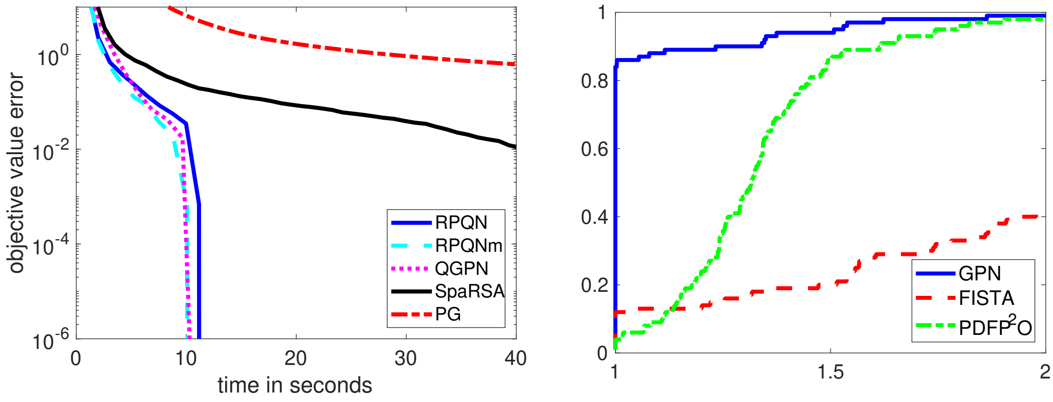
$$\varphi(x) = \lambda\|x\|_{2,1} = \lambda \sum_{j=1}^{p} \mu_j \|x_{[\mathcal{I}_j]}\|$$

for some $\lambda > 0$ and $\mu_j > 0$ for $j = 1, \dots, p$, which was already discussed in (6.7) in a similar form. In contrast, here the index sets $\mathcal{I}_j$ $(j = 1, \dots, p)$ are not required to be pairwise disjoint, i.e. the sets overlap, which explains the term *overlapping group regularizer*. In this case, to the author's knowledge no explicit formula for the proximity operator is known. On the other hand, we can write $\varphi = \tilde{\varphi} \circ B$, where $B$ is a linear mapping and $\tilde{\varphi}$ is a group penalty without overlapping. Thus, we can compute the proximity operator of $\tilde{\varphi}$, see Section 6.4.

Like in Section 6.3 we consider a logistic regression problem

$$\min_x \frac{1}{m} \sum_{i=1}^{m} \phi\big((Ax)_i\big) + \lambda \sum_{j=1}^{s} \mu_j \|x_{G_j}\|_2, \tag{6.11}$$

where $A \in \mathbb{R}^{m\times n}$ contains the information on feature vectors and corresponding labels and $\phi : \mathbb{R} \to \mathbb{R}$ is defined via $\phi(u) := \log\big(1 + \exp(-u)\big)$. A group penalty makes sense in many applications here, since some features are related to others and relations do not need to be disjoined. For more information on logistic regression with group penalty, we refer to [109].



(a) Objective value error related to the computation time for the methods tested in Section 6.6

(b) Performance profile showing the runtime for 100 random test examples from Section 6.7 with tolerance $10^{-6}$

Figure 6.11: Performance of the methods tested in Sections 6.6 and 6.7

**Algorithmic Details** As there is no formula to compute the proximity operator of $\varphi$, the subproblem solvers of the previous sections are not directly applicable. With the discussion above, both, the proximal Newton subproblem as well as the minimization problem to compute the proximity operator of $\varphi$, can be written in the form

$$\min_x \frac{1}{2} x^T Q x + c^T x + \tilde{\varphi}(Bx)$$

with $Q \in \mathbb{S}^n$ and $c \in \mathbb{R}^n$. We solve these problems with fixed point methods described by Chen et al. in [43]. For the computation of the proximity operator, we use the **F**ixed **P**oint algorithm based on the **P**roximity **O**perator (FP$^2$O) and for solving the proximal Newton subproblems the **P**rimal-**D**ual **F**ixed **P**oint algorithm based on the **P**roximity **O**perator (PDFP$^2$O) described in Section 6.1.1.
For both methods, we use a stopping tolerance of $10^{-9}$ and apply at most 10 iterations for each problem. The methods FP$^2$O and PDFP$^2$O require the largest eigenvalue of $BB^T$ (or an upper bound), which can be shown to be equal to the largest integer $k$ such that there exists an index $i \in \{1, \dots, n\}$ that is contained in $k$ index sets $\mathcal{I}_j$.
The computation of proximity operators is quite expensive here. Since the regularized proximal Newton-type method in Algorithm 5.1 requires an extra proximity operator for the criterion in (5.3) and the previous experiments have shown that the number of proximity evaluations is relatively high for this method, this method seems ineffective for the current example. Furthermore, as mentioned above, semismooth Newton methods are not applicable to this class of problems. Hence, we focus on the globalized proximal Newton method GPN and compare the results with FISTA. For the computation of the proximity operators, we also use FP$^2$O. Furthermore, we apply PDFP$^2$O directly to problem (6.11).

**Numerical Comparison** We follow an example in [6] and generate $A \in \mathbb{R}^{n \times m}$ with $n = 1000$, $m = 700$ and take its entries from a uniform distribution. The final matrix $A$ is then obtained by normalizing the columns of $A$. The groups $\mathcal{I}_j$ are

$$\begin{aligned}&\{1, \dots, 5\},\ \{5, \dots, 9\},\ \{9 \dots, 13\},\ \{13, \dots, 17\},\ \{17, \dots, 21\},\\ &\{4, 22, \dots, 30\},\ \{8, 31, \dots, 40\},\ \{12, 41, \dots, 50\},\ \{16, 51, \dots, 60\},\ \{20, 61, \dots, 70\},\\ &\{71, \dots, 80\},\ \{81, \dots, 90\},\ \dots,\ \{991, \dots 1000\}.\end{aligned}$$

The first five groups contain five consecutive numbers and the last element of one group is, at the same time, the first element of the next group. Each of the next five groups contain one element of one of the first groups. The remaining groups have no overlap and contain always 10 elements. Furthermore, the coefficients $\mu_j$ are chosen to be $1/\sqrt{|\mathcal{I}_j|}$, where $|\mathcal{I}_j|$ is the number of indices in that group.
The parameter $\lambda$ is chosen as $0.1\lambda_{\max}$, where $\lambda_{\max}$ is again the critical value such that zero is a solution of (6.11) for all $\lambda \geq \lambda_{\max}$. Let $a_i^T$ be the rows of $A$. Then a short computation shows

$$\lambda_{\max} = \frac{\sqrt{5}}{2m} \left\| \sum_{i=1}^m a_i \right\|_2 .$$

As before, we start with the initial value $x^0 = 0$.
We terminate each of the algorithms as soon as the current iterate satisfies (6.2) for `tol` $= 10^{-6}$. Again, we document the results using the performance profiles introduced by Dolan and Moré [57] on the runtime of 100 test examples. The results are shown in Figure 6.11(b), the averaged values for some counters are given in Table 6.7.

| method | iter | Newton- | sub- | matrix-vector |
|---|---|---|---|---|
| | | iter | iter | products |
| GPN | 9.5 | 9.5 | 95.1 | 221 |
| PDFP$^2$O | 76.9 | - | - | 156 |
| FISTA | 23.4 | - | 234 | 119 |

Table 6.7: Averaged values of 100 runs for the example in Section 6.7 using the tolerance $10^{-6}$ and three different methods.

We see that there are about 15% of the examples, where FISTA performs better than GPN, but in most examples GPN shows by far the best performance. This can be seen by looking at the number of inner iterations of both methods. Here, the costs of inner iterations is almost equal for both methods. Since the average number of inner iterations in FISTA is more then twice as big as the one of GPN, this illustrates the difference in performance.

At the end of this chapter it remains to briefly summarize the observations of our extensive analysis. It was shown that the proximal Newton-type algorithms developed in Chapters 4 and 5 are in some tests significantly superior to other state-of-the-art methods. Overall, no clear difference between the two methods can be identified. Instead, a decision for one of these methods depends on the chosen test setting. While the variants of the globalized proximal Newton-type method convinced in almost all test examples, for the regularized proximal quasi-Newton methods further work must be done on how quasi-Newton matrices can be appropriately updated if (6.1) does not hold. Since the number of proximity evaluations for RPQN are relatively big compared to QGPN, applying QGPN should be preferred, if the computation of the proximity operator is expensive.

Further, we note that according to our numerical results it might seem that the proximal gradient method from Algorithm 3.1 is no appropriate method for solving composite optimization problems. However, as mentioned at the beginning of this chapter, the purpose of the chapter was essentially to investigate the performance of our second-order proximal methods. The strength of the proximal gradient method, on the other hand, is usually shown in problems of much larger dimension, where using second-order information is too expensive.

# Chapter 7

# Comments and Outlook

In this thesis we investigated several proximal-type methods in order to find solutions of the composite optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) + \varphi(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function and $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex, proper and lower semicontinuous. In this chapter, we conclude the thesis by summarizing the main results and discuss some possible topics of future research.

### Proximal Gradient Method

After we collected a substantial amount of background material in Chapter 2, the formal study of proximal-type methods began in Chapter 3 with the introduction of the proximity operator. The theoretical properties of the proximity operator are well-understood, while its applications are almost endless for composite optimization problems. One basic and therefore important utilization is the proximal gradient method presented in Algorithm 3.1. Even though this method with line search strategy is not new (especially Tseng and Yun with their fundamental research [155] have to be mentioned), it seemed worthwhile to us to take a closer look at the method. On the one hand, it is fundamental for the subsequent investigated methods, on the other hand, some known convergence results have been collected and supplemented to obtain this collection. The main results are probably Lemma 3.11, which lead to the global convergence theorem, and the $\mathcal{O}(1/k)$-convergence of the sequence under a convexity assumption in Theorem 3.17.

There are multiple aspects of this theory which could lead to interesting developments or future research topics. One of the most obvious ideas would be to find a way to apply an acceleration scheme to the algorithm, thereby obtaining both, better theoretical convergence rates and better numerical performance. First-order proximal methods have become more and more attractive in recent years, especially for large scale problems, as in this case second-order information is generally expensive or not available. Since the evaluation of the proximity operator is usually more costly than a function evaluation, this development could be an important step forward and make an accelerated proximal gradient method with line search much more efficient than FISTA [14] and related methods currently are.

Although these were not studied extensively in Algorithm 3.1, it is possible to consider the matrices $H_k$ in more detail as well. It does not seem far-fetched that the speed of convergence can be influenced by a suitable choice of these matrices. In particular, approaches using the techniques of Barzilai and Borwein [11] could be useful here, see e.g. [164].

Finally, another modification is possible by weakening the assumptions: there are approaches to proximal-type methods for composite functions where $\varphi$ is not convex, but only lower semicontinuous [34, 35, 64, 83]. A reflection on the presented method and the study of the convergence properties under this condition seems quite attractive in view of the current research.

### Solving Proximal Subproblems

We continued the analysis in Section 3.3 by considering the subproblems that arise in all proximal-type methods whenever $H_k$ is no longer a multiple of the identity matrix. In doing so, we presented the details of an algorithm for solving these problems that combined the outstanding results from [16] with the compact representation of limited memory quasi-Newton matrices. Numerical experiments in Section 6 showed the superiority of this method over other approaches.

Even though the numerical analysis showed that the problems can usually be solved within a few steps with very high accuracy, a more detailed analysis regarding inexact solutions suggests itself here. In particular, this should include an investigation of how an error in the solution of the solved system of equations affects the accuracy of the final solution of the subproblem. Of course, this only makes sense if at the same time the inexactness conditions of the superordinate methods are adapted to it.

### Globalized Inexact Proximal Newton-type Method

The globalized inexact proximal Newton-type method was presented in Chapter 4 and combines an inexact proximal Newton-type method with the proximal gradient method from Section 3.2. This simple but ingenious combination based on a novel descent condition made it possible to prove far-reaching convergence results. These include the main global convergence results in Theorem 4.3 to prove that any accumulation point is stationary and Theorem 4.4, where suitable assumptions show the convergence of the complete sequence. Convergence rates were obtained using the Dennis-Moré condition for Kurdyka-Łojasiewicz functions in Theorem 4.8 and enhanced for strongly convex functions in Theorem 4.13.

Although the convergence theory appears very comprehensive and exhaustive, open questions remain for future research: The local theory is only valid under the condition of an isolated stationary point of the objective function. Since this is not fulfilled in some cases of interest, the theory should be extended to non-isolated stationary points. According to the author's considerations, this might be possible using a suitable error bound, similar to Assumption 5.11, or the uniformized Kurdyka-Łojasiewicz property [24].

Regarding first-order methods, a widely used criterion to obtain convergence in a nonconvex setting is a nonmonotone line search condition, see e.g. [83, 164]. It might therefore be worthwhile to investigate in Algorithm 4.1 in combination with such a criterion, especially when considering nonconvex problems.

Furthermore, we note that the globalization strategy from Algorithm 4.1 can be used to globalize several other locally convergent proximal Newton-type methods as [26, 95], thus improving their convergence behavior both theoretically and numerically. A closer look at the respective details of the algorithms should shed light on when this might be advantageous.

### Regularized Proximal Quasi-Newton Method

In Chapter 5 we presented a regularized proximal Newton-type method. In contrast to the previous methods, no classical step size search of Armijo-type was necessary to obtain convergence. Instead, a regularization parameter similar to trust-region methods was introduced. As before, comprehensive convergence results were proved under appropriate assumptions. For global convergence, these were Theorems 5.6 and 5.7. The error bound in Assumption 5.11 further allowed to show that any sequence of iterates has finite length (Theorem 5.13), and the Dennis-Moré condition provided the local convergence theorem including convergence rates (Theorem 5.15). A variant of the algorithm was added to the comprehensive theory in which it was combined with the proximal gradient method from Chapter 3. With the same theoretical properties as the original algorithm, the numerical results have thereby been improved significantly in some examples.

A major drawback of the method is that the global convergence theory only holds for real-valued convex function $\varphi$. As it seems reasonable that the results hold also for extended real-valued functions, this topic should be addressed theoretically in future.

It was further mentioned that functions with the introduced error bound condition satisfy the Kurdyka-Łojasiewicz property. Since the opposite does not need to hold, it might be a topic for future research to adapt the local convergence theory of the regularized proximal Newton-type method to hold for KL-functions.

Finally, in contrast to the globalized proximal Newton-type method, only the exact solution of the occurring subproblems was considered here. This was confirmed in Chapter 6 by the fact that the algorithms for the solution of these subproblems converge very fast and with very high accuracy. Nevertheless, the convergence theory with regard to the inexact solution of the subproblems (under suitable conditions to the inexactness) is another possible topic for future research.

The theoretical analysis was completed by an extensive numerical analysis of the introduced methods and in comparison to some state-of-the-art methods, where our focus was on proximal methods. The results show, that our methods perform very well in practical applications, and therefore verify the theoretical results.

## Final Comments

Structured optimization problems with composite functions arise in a wide range of application fields and their importance continues to increase, for example in the areas of machine learning and signal processing. Proximal methods in particular are used in this field. While most such methods are either globally convergent or have good local convergence properties, one of the driving factors which eventually led to the development of this thesis was to investigate algorithms that combine both. The material of the previous chapters underlines that in this context the detailed theoretical investigation of the developed methods and the numerical performance should not be treated separately. With this in mind, it is the author's hope that the theory, practical results and remarks presented throughout this thesis will prove useful to other researchers.

# Appendix A

# Additional Material

## A.1 The Proximal Gradient Method: Linear Convergence under Strong Convexity

In this section, we provide details of some results on proximity operators and the proximal gradient method. In detail, we first establish a result regarding the proximity operator with respect to different matrices. The subsequent analysis results in the proof of the convergence rate of the proximal gradient method under the assumption of $f$ being strongly convex, cf. Section 3.2.3 and in particular Theorem 3.18. This analysis is taken from Tseng and Yun [155, 156], where most of the stated results are shown in a more general setting. Due to their importance for our analysis, however, it seems worthwhile to provide the simplified analysis here.

Recall that we consider the optimization problem (1.1) with a convex, lower semicontinuous and proper function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$, and a function $f : \mathbb{R}^n \to \mathbb{R}$, which is continuously differentiable in a neighbourhood of $\operatorname{dom} \varphi$. For $H \in \mathbb{S}^n_{++}$ and $x \in \mathbb{R}^n$ let

$$r_H(x) := \operatorname{prox}^H_\varphi \left( x - H^{-1} \nabla f(x) \right) - x = \operatorname*{arg\,min}_{d \in \mathbb{R}^n} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + \varphi(x+d) \right\}.$$

Note that $r_{H_k}(x^k)$ coincides with the search direction $d^k$ in the proximal gradient method in Algorithm 3.1 and the second equality follows directly from the definition of the proximity operator (3.1). With this notation we present the proof of Lemma 3.6, which is a modification of [156, Lemma 3] and highly relevant for our analysis.

**Lemma A.1.** *Let $x \in \mathbb{R}^n$ and $H_1, H_2 \in \mathbb{S}^n_{++}$ be given. Then*

$$\|r_{H_2}(x)\| \leq \left( 1 + \frac{\lambda_{\max}(H_2)}{\lambda_{\min}(H_1)} \right) \cdot \frac{\lambda_{\max}(H_1)}{\lambda_{\min}(H_2)} \cdot \|r_{H_1}(x)\|.$$

*Proof.* Set $d^1 := r_{H_1}(x)$ and $d^2 := r_{H_2}(x)$ and define the mapping $\overline{\varphi} : d \mapsto \varphi(x+d)$. Then, $s \in \partial \overline{\varphi}(d)$ if and only if the subdifferential inequality

$$\varphi(x+y) \geq \varphi(x+d) + s^T(y-d)$$

holds for all $y \in \mathbb{R}^n$. Substituting $z := x + y$, this yields $s \in \partial \varphi(x+d)$ and we get $\partial \overline{\varphi}(d) = \partial \varphi(x+d)$. Using Fermat's rule (Proposition 2.47) and (3.1), we get

$$0 \in \nabla f(x) + H_1 d^1 + \partial \varphi(x + d^1).$$

The reverse implications of these results yield

$$d^1 \in \operatorname*{arg\,min}_{d\in\mathbb{R}^n} \left\{(\nabla f(x) + H_1 d^1)^T d + \varphi(x+d)\right\}.$$

Thus, by the subdifferential inequality, we obtain

$$(\nabla f(x) + H_1 d^1)^T d^1 + \varphi(x+d^1) \le (\nabla f(x) + H_1 d^1)^T d^2 + \varphi(x+d^2), \tag{A.1}$$

and, applying the same arguments to $d^2$,

$$(\nabla f(x) + H_2 d^2)^T d^2 + \varphi(x+d^2) \le (\nabla f(x) + H_2 d^2)^T d^1 + \varphi(x+d^1).$$

Adding up both inequalities yields

$$(d^1)^T H_1 d^1 - (d^1)^T (H_1 + H_2) d^2 + (d^2)^T H_2 d^2 \le 0,$$

on which is equivalent to

$$\left\|H_1^{1/2} d^1 + \tfrac{1}{2} H_1^{-1/2}(H_1+H_2)d^2\right\|^2 - \frac{1}{4}\left\|H_1^{-1/2}(H_1+H_2)d^2\right\|^2 + (d^2)^T H_2 d^2 \le 0,$$

where $H_1^{1/2}$ denotes the matrix square root of $H_1$. We substitute $u^1 = H_1^{1/2} d^1$, $u^2 = H_1^{1/2} d^2$ and $Q = H_1^{-1/2} H_2 H_1^{-1/2}$ to rewrite this as

$$\left\|u^1 - \tfrac{1}{2}(I+Q)u^2\right\|^2 \le \frac{1}{4}\|(I+Q)u^2\|^2 - (u^2)^T Q u^2 = \frac{1}{4}\|(I-Q)u^2\|^2.$$

Taking the square root on both sides and using the triangle inequality to the left hand side yields

$$\frac{1}{2}\|(I+Q)u^2\| - \|u^1\| \le \left\|u^1 - \tfrac{1}{2}(I+Q)u^2\right\| \le \frac{1}{2}\|(I-Q)u^2\|,$$

which gives the estimate

$$\frac{1}{2}\|(I+Q)u^2\| - \frac{1}{2}\|(I-Q)u^2\| \le \|u^1\|.$$

By multiplication of both sides by $2\|(I+Q)u^2\| + 2\|(I-Q)u^2\|$, we obtain

$$4(u^2)^T Q u^2 \le 2\|u^1\|\big(\|(I+Q)u^2\| + \|(I-Q)u^2\|\big) \le 4\|u^1\| \cdot (1 + \lambda_{\max}(Q))\|u^2\|.$$

Next, we use the estimates

$$(u^2)^T Q u^2 = (d^2)^T H_2 d^2 \ge \lambda_{\min}(H_2)\|d^2\|^2 \qquad \text{and} \qquad \|u^i\| \le \sqrt{\lambda_{\max}(H_1)}\|d^i\|$$

for $i = 1, 2$ and get

$$\begin{aligned}\|d^2\|^2 &\le \frac{(d^2)^T H_2 d^2}{\lambda_{\min}(H_2)} \\ &= \frac{(u^2)^T Q u^2}{\lambda_{\min}(H_2)} \\ &\le \frac{\|u^1\|(1+\lambda_{\max}(Q))\|u^2\|}{\lambda_{\min}(H_2)}\end{aligned}$$

$$\leq (1 + \lambda_{\max}(Q)) \frac{\lambda_{\max}(H_1)}{\lambda_{\min}(H_2)} \|d^1\| \cdot \|d^2\|.$$

The claim follows by division through $\|d^2\|$ (note that the estimate in the theorem is trivial for $\|d^2\| = 0$) and

$$\begin{aligned} \lambda_{\max}(Q) &= \max_{y \neq 0} \frac{x^T H_1^{-1/2} H_2 H_1^{-1/2} y}{y^T y} = \max_{z \neq 0} \frac{z^T H_2 z}{z^T H_1 z} = \max_{z \neq 0} \left( \frac{z^T H_2 z}{z} \frac{z^T z}{z^T H_1 z} \right) \\ &\leq \max_{z \neq 0} \frac{z^T H_2 z}{z} \cdot \max_{z \neq 0} \frac{z^T z}{z^T H_1 z} = \frac{\lambda_{\max}(H_2)}{\lambda_{\min}(H_1)}. \end{aligned}$$

□

The following results are auxiliary results for the proof of the convergence rate for the proximal gradient method. The proofs are taken from Lemma 5(a) and Theorem 4 in [156]. For these results we use the notation from Algorithm 3.1.

**Lemma A.2.** *With the notation from Algorithm 3.1 there holds for any $y \in \mathbb{R}^n$ and $k \geq 0$*

$$(\nabla f(x^k) + H_k d^k)^T (x^{k+1} - y) + \varphi(x^{k+1}) - \varphi(y) \leq (t_k - 1)\big((d^k)^T H_k d^k + \Delta_k\big).$$

*Proof.* Analogous to (A.1) we get

$$(\nabla f(x^k) + H_k d^k)^T d^k + \varphi(x^k + d^k) \leq (\nabla f(x^k) + H_k d^k)^T (y - x^k) + \varphi(y)$$

for any $y \in \mathbb{R}^n$. Since $x^{k+1} = x^k + t_k d^k$, we have

$$\begin{aligned} &(\nabla f(x^k) + H_k d^k)^T (x^{k+1} - y) + \varphi(x^{k+1}) - \varphi(y) \\ =&(t_k - 1)(\nabla f(x^k) + H_k d^k)^T d + \varphi(x^{k+1}) + (\nabla f(x^k) + H_k d^k)^T (x^k + d^k - y) - \varphi(y) \\ \leq&(t_k - 1)(\nabla f(x^k) + H_k d^k)^T d^k + \varphi(x^{k+1}) - \varphi(x^k + d^k) \\ =&(t_k - 1)(\nabla f(x^k) + H_k d^k)^T d^k + \varphi(x^k + t_k d^k) - \varphi(x^k + d^k) \\ \leq&(t_k - 1)(\nabla f(x^k) + H_k d^k)^T d^k + (1 - t_k)\varphi(x^k) + t_k \varphi(x^k + d^k) - \varphi(x^k + d^{)} \\ =&(t_k - 1)\big(\nabla f(x^k)^T d^k + (d^k)^T H_k d^k + \varphi(x^k + d^k) - \varphi(x^k)\big) \\ =&(t_k - 1)\big((d^k)^T H_k d^k + \Delta_k\big), \end{aligned}$$

where we use the convexity of $\varphi$ in the fourth step. □

**Lemma A.3.** *Assume that $f$ is strongly convex and $\nabla f$ is Lipschitz continuous. Let $x^*$ be the unique minimizer of $\psi$. If $mI \preceq H_k \preceq MI$ holds for $0 < m \leq M$, then there exists $\tau > 0$ (independently of $k$) such that*

$$\|x^k - x^*\| \leq \tau \|d^k\|$$

*holds for all $k \geq 0$.*

*Proof.* Since $f$ is strongly convex, $\psi$ is strongly coercive and, hence, has a unique minimizer $x^*$, cf. Corollary 2.17. Similar to (A.1) with $H_1 = I$ and $x + d^2 = x^*$ we have

$$(\nabla f(x^k)^T + r_I(x^k))^T r_I(x^k) + \varphi(x^k + r_I(x^k)) \leq (\nabla f(x^k) + r_I(x^k))^T (x^* - x^k) + \varphi(x^*)$$

and with $x$ replaced by $x^*$ and $d^2 = r_I(x)$

$$\varphi(x^*) \leq \nabla f(x^*)^T(x^k + r_I(x^k) - x^*) + \varphi(x^k + r_I(x^k)),$$

since $d_I(x^*) = 0$ due to Proposition 3.7. Adding both inequalities and simplifying yields

$$(\nabla f(x^*) - \nabla f(x^k))^T(x^* - x^k) + \|r_I(x^k)\|^2 \leq (\nabla f(x^*) - \nabla f(x^k))^T r_I(x^k) + r_I(x^k)^T(x^* - x^k). \tag{A.2}$$

Since $f$ is strongly convex, its gradient is strongly monotone with modulus $\mu > 0$ (Proposition 2.22), hence

$$\mu\|x^* - x^k\|^2 \leq (\nabla f(x^*) - \nabla f(x^k))^T(x^* - x^k).$$

Using the Lipschitz continuity of $\nabla f$ with some Lipschitz constant $L \geq \mu$, we obtain from (A.2)

$$\mu\|x^k - x^*\|^2 + \|r_I(x^k)\|^2 \leq L\|x^k - x^*\| \; \|r_I(x^k)\| + \|x^k - x^*\| \; \|r_I(x^k)\|,$$

from which it finally follows that $\mu\|x^k - x^*\|^2 \leq (L + on1)\|x^k - x^*\| \; \|r_I(x^k)\|$. We divide both sides by $\mu\|x^k - x^*\|$ to obtain

$$\|x^k - x^*\| \leq \frac{L+1}{\mu}\|r_I(x^k)\|.$$

By Lemma A.1 with $H_2 = I$ and $H_1 = H_k$ we know

$$\|r_I(x^k)\| \leq \left(1 + \frac{1}{\lambda_{\min}(H_k)}\right) \cdot \lambda_{\max}(H_k)\|d^k\| \leq \frac{(1+m)M}{m}\|d^k\|.$$

Hence, the assertion holds with $\tau = \frac{L+1}{\mu} \cdot \frac{(1+m)M}{m}$. □

After these auxiliary results are available, we can prove Theorem 3.18, which provides the convergence rate of the proximal gradient method.

**Theorem A.4.** *Suppose that $f$ is strongly convex, $\nabla f$ is Lipschitz continuous, and $\psi$ is bounded from below with $x^* \in \arg\min \psi$. Furthermore let $0 < m \leq M$ such that $mI \preceq H_k \preceq MI$. Then Algorithm 3.1 generates a sequence $\{x^k\}$ that satisfies*

$$\psi(x^{k+1}) - \psi(x^*) \leq c_1\big(\psi(x^k) - \psi(x^*)\big),$$

*for all $k \geq 0$, where $c_1 \in (0,1)$ is a constant depending on the Lipschitz constant of $\nabla f$, the strong convexity modulus of $f$, $m$, $M$ and the constants $\sigma$, $\beta$.*

*Proof.* We follow the proof of Theorem 5.2 in [155]. For fixed $k \geq 0$ let $\zeta^k$ be a point on the straight line between $x^{k+1}$ and $x^*$ such that

$$\begin{aligned}
\psi(x^{k+1}) - \psi(x^*) &= f(x^{k+1}) + \varphi(x^{k+1}) - f(x^*) - \varphi(x^*) \\
&= \nabla f(\xi^k)^T(x^{k+1} - x^*) + \varphi(x^{k+1}) - \varphi(x^*) \\
&= \big(\nabla f(\xi^k) - \nabla f(x^k)\big)^T(x^{k+1} - x^*) - (H_k d^k)(x^{k+1} - x^*) \\
&\quad + (\nabla f(x^k) + H_k d^k)^T(x^{k+1} - x^*) + \varphi(x^{k+1}) - \varphi(x^*)
\end{aligned}$$

Applying the Lipschitz continuity of $\nabla f$ with some Lipschitz constant $L > 0$ and Lemma

A.2, this yields

$$\begin{aligned}&\psi(x^{k+1}) - \psi(x^*)\\ &\quad\leq L\|\xi^k - x^k\|\ \|x^{k+1} - x^*\| + \|H_k d^k\|\ \|x^{k+1} - x^*\| + (t_k - 1)\big((d^k)^T H_k d^k + \Delta_k\big). \quad (A.3)\end{aligned}$$

With the estimate $\|x^{k+1} - x^k\| = t_k\|d^k\| \leq \|d^k\|$ and Lemma A.3, we have

$$\begin{aligned}\|\xi^k - x^k\| &\leq \|x^{k+1} - x^k\| + \|x^k - x^*\| \leq (1+\tau)\|d^k\|\\ \|x^{k+1} - x^*\| &\leq \|x^{k+1} - x^k\| + \|x^k - x^*\| \leq (1+\tau)\|d^k\|.\end{aligned}$$

We plug this into (A.3) to obtain

$$\psi(x^{k+1}) - \psi(x^*) \leq \big[L(1+\tau)^2 + M(1+\tau) + M\big]\|d^k\|^2 + (t_k - 1)\Delta_k \leq -c_2\Delta_k$$

with

$$c_2 = \big[L(1+\tau)^2 + M(1+\tau) + M\big]/m + 1 - t_{\min} > 0,$$

where we used Lemma 3.9 and $t_k \geq t_{\min}$, cf. (3.8). Together with (3.5) this yields (note that $\Delta_k < 0$)

$$\psi(x^{k+1}) - \psi(x^*) \leq \frac{c_2}{\sigma t_{\min}}\big(\psi(x^k) - \psi(x^{k+1})\big),$$

which reformulates to

$$\psi(x^{k+1}) - \psi(x^*) \leq c_1\big(\psi(x^k) - \psi(x^*)\big)$$

with $c_1 = c_2/(c_2 + \sigma t_{\min})$. Hence, the linear convergence of $\{\psi(x^k)\}$ is shown. □

At the end of this section we note, that it is a short step from this result to prove R-linear convergence of the sequence $\{x^k\}$, which means that

$$\limsup_{k\to\infty} \|x^k - x^*\|^{1/k} < 1.$$

We skip the proof and refer to [155] for details.

## A.2 Convergence Rates of GIPN under the KL-Property

This section deals with the proof of the convergence rates in Theorem 4.8 for the globalized inexact proximal Newton-type method in Algorithm 4.1. For this purpose, we use the notation and preliminaries of Chapter 4, especially in Section 4.3. The analysis is developed by Bonettini et al. in [26] for VMILAn, which is an algorithm with some similarities to the presented framework in section 4.1. Hence, the following results mainly coincide with [26]. Furthermore, we focus on the differences to this work and skip some of the details, in case they are identical in our analysis and in [26].

In the subsequent analysis, we assume that the premises of Theorem 4.7 hold. In particular, $\{H_k\}$ satisfies the Dennis-Moré condition (4.13) and is uniformly bounded and positive definite, i.e. $mI \preceq H_k \preceq MI$ holds for all $k \geq 0$ with suitable $0 < m \leq M$, and $x^*$ is an accumulation point of a sequence $\{x^k\}$ generated by Algorithm 4.1, which is an isolated stationary point of $\psi$ and satisfies the KL-property.

In view of Theorem 4.7, there exists $k_0 \geq 0$, such that for all $k \geq k_0$, the search direction is attained by the inexact proximal Newton-type direction and the full step size $t_k = 1$ is

accepted. For simplicity in the subsequent analysis, we assume without loss of generality $k_0 = 0$. In particular, this means $x^{k+1} = x^k + d^k$ for all $k \geq 0$.
Note that the above assumptions imply that the complete sequence $\{x^k\}$ converges to $x^*$ and is therefore bounded. Hence there is a compact, convex set, containing this sequence such that $\nabla f$ is Lipschitz continuous with Lipschitz constant $L > 0$ on this set.
We start with some simple observations, which result from the definitions and some previous results.

**Lemma A.5.** *For all $k \geq 0$ the following estimates hold:*
*(a) $\psi(x^{k+1}) \leq \psi(x^k) - a\|x^{k+1} - x^k\|^2$ for some $a > 0$,*
*(b) $0 \leq -\sum_{k=0}^{\infty} \Delta_k < +\infty$.*

*Proof.* For part (a) we combine the Armijo-type line search (4.8) with Lemma 3.9 to get

$$\psi(x^{k+1}) - \psi(x^k) \leq \sigma\Delta_k \leq -\sigma m\|d_k\|^2 = -\sigma m\|x^{k+1} - x^k\|^2,$$

noting that we have $t_k = 1$. Furthermore, since $\{x^k\}$ converges to $x^*$, we get

$$+\infty > \psi(x^0) - \psi(x^*) = \sum_{k=0}^{\infty} \psi(x^k) - \psi(x^{k+1}) \geq -\sigma \sum_{k=0}^{\infty} \Delta_k \geq 0,$$

where the last inequality comes from $\Delta_k \leq 0$, cf. Lemma 3.9. This yields the claim of (b). □

on The next result has the purpose to deduce the assumption in (4.15) on the size of a subgradient of $\psi$ in the new iterate $x^{k+1}$. For that purpose, we introduce the *$\varepsilon$-subdifferential* $\partial_\varepsilon\psi(x)$, which is the set of all *$\varepsilon$-subgradients* $s \in \mathbb{R}^n$ satisfying

$$\psi(y) \geq \psi(x) + s^T(y - x) - \varepsilon.$$

Note that, in particular, $\partial_0\psi(x) = \partial\psi(x)$. For more information about the $\varepsilon$- or approximate subdifferential and its properties, we refer to the monograph [76]. With this definition we give the following estimate.

**Lemma A.6.** *There exist $\bar{\varepsilon}_k, \hat{\varepsilon}_k \geq 0$ with $\bar{\varepsilon}_k + \hat{\varepsilon}_k \leq \varepsilon_k$, and an approximate subgradient $s^k \in \nabla f(x^{k+1}) + \partial_{\bar{\varepsilon}_k}\varphi(x^{k+1})$ such that*

$$\|s^k\| \leq \alpha_1\|d^k\| + \nu_k,$$

*where $\alpha_1 > 0$, $\varepsilon_k = q_k(d^k) - q_k(d^k_{ex}) \geq 0$ and $\nu_k = \mathcal{O}(\sqrt{\varepsilon_k})$.*

*Proof.* Note that $d^k_{ex}$ denotes the exact minimizer of $q_k$. Thus, we know $0 \in \partial q_k(d^k_{ex})$ using Fermat's theorem. Let $\varepsilon_k := q_k(d^k) - q_k(d^k_{ex}) \geq 0$. Then, we also have $0 \in \partial_{\varepsilon_k} q_k(d^k)$. Using Theorem 3.1.1 and Example 1.2.2 in [76], we get

$$\begin{aligned}
\partial_{\varepsilon_k}(d^k) &= \bigcup_{0 \leq \bar{\varepsilon}_k + \hat{\varepsilon}_k \leq \varepsilon_k} \partial_{\hat{\varepsilon}_k}\big(\nabla f(x^k)^T \cdot + \tfrac{1}{2} \cdot^T H_k \cdot\big)(d^k) + \partial_{\bar{\varepsilon}_k}\varphi(x^{k+1}) \\
&= \bigcup_{0 \leq \bar{\varepsilon}_k + \hat{\varepsilon}_k \leq \varepsilon_k} \big\{\nabla f(x^k) + H_k d^k + H_k e : \tfrac{1}{2} e^T H_k e \leq \hat{\varepsilon}_k\big\} + \partial_{\bar{\varepsilon}_k}\varphi(x^{k+1}).
\end{aligned}$$

Thus, there exist $\bar{\varepsilon}_k, \hat{\varepsilon}_k \geq 0$ such that $\bar{\varepsilon}_k + \hat{\varepsilon}_k \leq \varepsilon_k$ and

$$0 = \nabla f(x^k) + H_k d^k + H_k e + w^k,$$

where $\frac{1}{2}e^T H_k e \le \hat{\varepsilon}_k$ and $w^k \in \partial_{\hat{\varepsilon}_k}\varphi(x^{k+1})$. Set $s^k = \nabla f(x^{k+1}) + w^k$. Then

$$\begin{aligned}\|s^k\| &= \big\|\nabla f(x^{k+1}) - \nabla f(x^k) - H_k d^k - H_k e\big\| \\ &\le \big\|\nabla f(x^{k+1}) - \nabla f(x^k)\| + \|H_k\| \cdot \|d^k\| + \|H_k e\| \\ &\le (L+M)\|d^k\| + \sqrt{\frac{2}{m}} M \sqrt{\hat{\varepsilon}_k},\end{aligned}$$

where we used the estimate

$$\|H_k e\|^2 \le M^2 \|e\|^2 \le \frac{M^2}{m}\|e\|_{H_k}^2 \le \frac{M^2}{m}\hat{\varepsilon}_k$$

for the final inequality. This yields the claim with $\alpha_1 := L + M$ and $\nu_k = \sqrt{\frac{2}{m}} M \sqrt{\varepsilon_k}$. □

The above statement does only make sense if we know that $\{\varepsilon_k\}$ is a vanishing sequence. This follows from the (local) Lipschitz continuity of $q_k$ (noting that we can state an upper bound on the Lipschitz constant independently of $k$) in combination with Proposition 4.5(a) noting that $r(x^k) \to 0$ for $k \to \infty$. Nevertheless, the assumption used in [26] for the following results is little stronger than the one shown in Lemma A.6. In detail, we need $\nu_k = \mathcal{O}(|\Delta_k|)$, whereas the above only yields $\nu_k^2 = \mathcal{O}(|\Delta_k|)$. Details on the deduction of this estimate are left to the reader.
Furthermore, [26, Theorem 1] proves that under this assumption the sequence $\{x^k\}$ has finite length, i.e. $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| < \infty$. We continue our analysis with some technical estimates, which lead to the final convergence proof.

**Lemma A.7.** *In addition to the above assumptions assume that the sequence $\{x^k\}$ satisfies the following condition: For every $k \ge 0$ there exists $s^k \in \partial\psi(x^k + d^k)$ such that*

$$\|s^k\| \le \alpha_1 \|d^k\| + \alpha_2 |\Delta_k| \tag{A.4}$$

*holds for some $\alpha_1, \alpha_2 \ge 0$. Furthermore, assume that the KL-property is satisfied in $x^*$ with the function $\phi(s)$. Then the following estimates hold for sufficiently large $k \ge 0$.*

*(a)* $2\|x^{k+1} - x^k\| \le \|x^k - x^{k-1}\| + \phi_k - \frac{\alpha_2}{\alpha_1}\Delta_k,$
*where* $\phi_k = \frac{\alpha_1}{a}\big(\phi(\psi(x^k) - \psi(x^*)) - \phi(\psi(x^{k+1}) - \psi(x^*))\big),$

*(b)* $\|x^k - x^*\| \le \Big(\frac{1}{\sqrt{a}} + \frac{\alpha_1}{a} + \frac{\alpha_2}{\alpha_1 \sigma}\Big)\overline{\phi}\big(\psi(x^{k-1}) - \psi(x^*)\big),$
*where* $\overline{\phi}(t) := \max\{\phi(t), \sqrt{t}\}$.

*Proof.* (a) For sufficiently large $k \ge 0$, we have $x^k \in U \cap [\psi(x^*) < \psi < \psi(x^*) + \nu]$, where $\phi, \nu$ and $U$ come from the definition of the KL-property. The KL-inequality in $x^k$ and (A.4) yield

$$\phi'\big(\psi(x^k) - \psi(x^*)\big) \ge \frac{1}{\|s^{k-1}\|} \ge \frac{1}{\alpha_1\|d^{k-1}\| + \alpha_2|\Delta_k|}, \tag{A.5}$$

where we note that the denominators do not vanish due to the KL-inequality. Since $\phi$ is concave, we use (A.5) and Lemma A.5 to obtain

$$\begin{aligned}\phi\big(\psi(x^k) - \psi(x^*)\big) - \phi\big(\psi(x^{k+1}) - \psi(x^*)\big) &\ge \phi'\big(\psi(x^k) - \psi(x^*)\big)\big(\psi(x^k) - \psi(x^{k+1})\big) \\ &\ge \frac{a\|d^k\|^2}{\alpha_1\|d^{k-1}\| + \alpha_2|\Delta_k|}.\end{aligned}$$

Rearranging terms and using $d^k = x^{k+1} - x^k$ yields

$$\|x^{k+1} - x^k\|^2 \leq \phi_k \frac{\alpha_1 \|d^{k-1}\| + \alpha_o n 2|\Delta_k|}{\alpha_1}.$$

The claim of (a) follows from taking the square root and applying the inequality $2\sqrt{uv} \leq u+v$.

(b) Let $k_0 \geq 0$ such that (a) holds for all $k \geq k_0$ and $N > k_0$. Summing the inequality in part (a) for $k = k_0, \ldots, N$ yields

$$2\sum_{k=k_0}^{N} \|x^{k+1} - x^k\| \leq \sum_{k=k_0}^{N} \|x^k - x^{k-1}\| + \sum_{k=k_0}^{N} \phi_k - \frac{\alpha_2}{\alpha_1} \sum_{k=k_0}^{N} \Delta_k.$$

We rearrange terms and use the triangle inequality to obtain

$$\begin{aligned}
\|x^{k_0} - x^{N+1}\| &\leq \sum_{k=k_0}^{N} \|x^{k+1} - x^k\| \\
&\leq \|x^{k_0} - x^{k_0-1}\| - \|x^{N+1} - x^N\| \\
&\quad + \frac{\alpha_1}{a}\big(\phi(\psi(x^{k_0}) - \psi(x^*)) - \phi(\psi(x^{N+1}) - \psi(x^*))\big) - \frac{\alpha_2}{\alpha_1} \sum_{k=k_0}^{N} \Delta_{k-1} \\
&\leq \|x^{k_0} - x^{k_0-1}\| + \frac{\alpha_1}{a}\phi(\psi(x^{k_0}) - \psi(x^*)) - \frac{\alpha_2}{\alpha_1} \sum_{k=k_0}^{N} \Delta_k \\
&\leq \frac{1}{\sqrt{a}}\sqrt{\psi(x^{k_0-1}) - \psi(x^{k_0})} + \frac{\alpha_1}{a}\phi(\psi(x^{k_0}) - \psi(x^*)) \\
&\quad + \frac{\alpha_2}{\alpha_1 \sigma}\big(\psi(x^{k_0}) - \psi(x^{N+1})\big),
\end{aligned}$$

where we used the line search (4.8) and Lemma A.5 (a) for the last estimate. We use $\psi(x^{k_0-1}) \geq \psi(x^{k_0}) \geq \psi(x^*)$, the monotonicity of $\phi$ and take the limit $N \to \infty$ to get

$$\|x^{k_0} - x^*\| \leq \frac{1}{\sqrt{a}}\sqrt{\psi(x^{k_0-1}) - \psi(x^*)} + \frac{\alpha_1}{a}\phi(\psi(x^{k_0-1}) - \psi(x^*)) + \frac{\alpha_2}{\alpha_1 \sigma}\big(\psi(x^{k_0-1}) - \psi(x^*)\big).$$

Finally, for sufficiently large $k_0 \geq 0$ we have $\psi(x^{k_0-1}) - \psi(x^*) \leq \sqrt{\psi(x^{k_0-1}) - \psi(x^*)}$, which completes the proof. □

The consequence of this preliminary work is the convergence rate theorem for Algorithm 4.1, which is stated in Theorem 4.8. The main part of the proof is the theory of Frankel et al. in [64].

**Theorem A.8.** *Let the assumptions of Lemma A.7 hold and assume that the KL-function in $x^*$ is $\phi(s) = \frac{C}{\theta} \cdot s^\theta$ for some $C > 0$ and $\theta \in (0, 1]$. Then the following hold:*
*(a) If $\theta = 1$, then $\{x^k\}$ converges in a finite number of steps.*
*(b) If $\theta \in [\frac{1}{2}, 1)$, there exists $\delta > 0$ such that*

$$\psi(x^k) - \psi(x^*) = \mathcal{O}(e^{-\delta k}), \quad \text{and} \quad \|x^k - x^*\| = \mathcal{O}(e^{-\delta/2} k).$$

*(c) If $\theta \in (0, \frac{1}{2})$, there exists $k_0 \geq 0$ such that*

$$\psi(x^k) - \psi(x^*) = \mathcal{O}\big((k-k_0)^{-\frac{1}{1-2\theta}}\big) \quad \textit{and} \quad \|x^k - x^*\| = \mathcal{O}\big((k-k_0+1)^{-\frac{1}{1-2\theta}}\big).$$

*Proof.* We know that the sequence $\{x^k\}$ converges to $x^*$ by Theorem 4.4, see also the discussion at the beginning of Section 4.3. Hence, there exists $k_0 \geq 0$ such that $x^k \in U \cap [\psi(x^*) < \psi < \psi(x^*) + \nu]$ holds for all $k \geq k_0$, where $\phi, \nu$ and $U$ come from the definition of the KL-property.

Using (4.8), we get

$$|\Delta_k| \leq \frac{1}{\sigma}\big(\psi(x^k) - \psi(x^{k+1})\big). \tag{A.6}$$

We use (A.4), take the squares of both sides and multiply by $a/\alpha_1^2$ to obtain

$$\begin{aligned}\frac{a}{\alpha_1^2}\|s^k\|^2 &\leq a\|x^{k+1} - x^k\|^2 + a\frac{\alpha_2^2}{\alpha_1^2}\Delta_k^2 + 2a\frac{\alpha_2}{\alpha_1}\Delta_k\|x^{k+1} - x^k\| \\ &\leq \big(\psi(x^k) - \psi(x^{k+1})\big) + a\frac{\alpha_2^2}{\alpha_1^2}\Delta_k^2 + 2\sqrt{a}\frac{\alpha_2}{\alpha_1}\Delta_k\sqrt{\psi(x^k) - \psi(x^{k+1})},\end{aligned}$$

where we again applied Lemma A.5(a).

Since $\lim_{k\to\infty} \Delta_k = 0$, see (A.6), we can choose $k_0$ sufficiently large such that $\Delta_k^2 \leq \Delta_k \leq \sqrt{\Delta_k}$ holds for all $k \geq k_0$. Together with (A.6) we get

$$\frac{a}{\alpha_1^2}\|s^k\|^2 \leq \mu\big(\psi(x^k) - \psi(x^{k+1})\big)$$

with $\mu = 1 + a\frac{\alpha_1^2}{\alpha_2^2\sigma} + 2\frac{\alpha_2}{\alpha_1}\sqrt{\frac{a}{\sigma}}$. Set $u_k := \psi(x^k) - \psi(x^*)$. By multiplying each side of the inequality by $\phi'(u_{k+1})$, we have

$$\mu\phi'(u_{k+1})^2(u_k - u_{k+1}) \geq \frac{a}{\alpha_1^2}\phi'(u_{k+1})^2\|s^k\|^2 \geq \frac{a}{\alpha_1^2},$$

where the second estimate follows from the KL-inequality. Following the proof of [26, Theorem 3], this is equivalent to equation (6) in [64, Theorem 3.4], from which the convergence rates for $\psi(x^k) - \psi(x^*)$ follow, whereas the ones for $\|x^k - x^*\|$ are obtained using Lemma A.7(b), and this completes the proof. □

# Bibliography

[1] G. Andrew and J. Gao, *Scalable training of $L_1$-regularized log-linear models*, in Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 33–40.

[2] N. Antonello, L. Stella, P. Patrinos, and T. van Waterschoot, *Proximal gradient algorithms: Applications in signal processing*, arXiv preprint arXiv:1803.01621, (2018).

[3] A. Aravkin, J. V. Burke, and G. Pillonetto, *Robust and trend-following Student's t Kalman smoothers*, SIAM Journal on Control and Optimization, 52 (2014), pp. 2891–2916.

[4] A. Aravkin, M. P. Friedlander, F. J. Herrmann, and T. Van Leeuwen, *Robust inversion, dimensionality reduction, and randomized sampling*, Mathematical Programming, 134 (2012), pp. 101–125.

[5] A. Y. Aravkin, R. Baraldi, and D. Orban, *A proximal quasi-Newton trust-region method for nonsmooth regularized optimization*, arXiv preprint arXiv:2103.15993, (2021).

[6] A. Argyriou, C. A. Micchelli, M. Pontil, L. Shen, and Y. Xu, *Efficient first order methods for linear composite regularizers*, arXiv preprint arXiv:1104.1436, (2011).

[7] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality*, Mathematics of Operations Research, 35 (2010), pp. 438–457.

[8] A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM Journal on Optimization, 16 (2006), pp. 697–725.

[9] G. Ayers and J. C. Dainty, *Iterative blind deconvolution method and its applications*, Optics Letters, 13 (1988), pp. 547–549.

[10] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Structured sparsity through convex optimization*, Statistical Science, 27 (2012), pp. 450–468.

[11] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA Journal of Numerical Analysis, 8 (1988), pp. 141–148.

[12] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer International Publishing, 2nd ed., 2017.

[13] A. Beck, *First-Order Methods in Optimization*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 2017.

[14] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[15] S. Becker and J. Fadili, *A quasi-Newton proximal splitting method*, in Advances in Neural Information Processing Systems, 2012, pp. 2618–2626.

[16] S. Becker, J. Fadili, and P. Ochs, *On quasi-Newton forward-backward splitting: Proximal calculus and convergence*, SIAM Journal on Optimization, 29 (2019), pp. 2445–2481.

[17] J. Y. Bello Cruz and T. T. Nghia, *On the convergence of the forward–backward splitting method with linesearches*, Optimization Methods and Software, 31 (2016), pp. 1209–1238.

[18] M. Benning, L. Gladden, D. Holland, C.-B. Schönlieb, and T. Valkonen, *Phase reconstruction from velocity-encoded MRI measurements–a survey of sparsity-promoting variational approaches*, Journal of Magnetic Resonance, 238 (2014), pp. 26–43.

[19] M. Bertero, D. Bindi, P. Boccacci, M. Cattaneo, C. Eva, and V. Lanza, *A novel blind-deconvolution method with an application to seismology*, Inverse Problems, 14 (1998), pp. 815–833.

[20] D. Bertsekas, *Nonlinear Programming*, Athena Scientific Optimization and Computation Series, Athena Scientific, 2016.

[21] J. Bolte, P. L. Combettes, and J.-C. Pesquet, *Alternating proximal algorithm for blind image recovery*, in 2010 IEEE International Conference on Image Processing, IEEE, 2010, pp. 1673–1676.

[22] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM Journal on Optimization, 18 (2007), pp. 556–572.

[23] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity*, Transactions of the American Mathematical Society, 362 (2010), pp. 3319–3363.

[24] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.

[25] S. Bonettini, I. Loris, F. Porta, and M. Prato, *Variable metric inexact line-search-based methods for nonsmooth optimization*, SIAM Journal on Optimization, 26 (2016), pp. 891–921.

[26] S. Bonettini, I. Loris, F. Porta, M. Prato, and S. Rebegoldi, *On the convergence of a linesearch based proximal-gradient method for nonconvex optimization*, Inverse Problems, 33 (2017), pp. 055005,30.

[27] S. Bonettini, P. Ochs, M. Prato, and S. Rebegoldi, *An abstract convergence framework with application to inertial inexact forward-backward methods*, tech. report, 2021.

[28] S. Bonettini, F. Porta, and V. Ruggiero, *A variable metric forward-backward method with extrapolation*, SIAM Journal on Scientific Computing, 38 (2016), pp. A2558–A2584.

[29] S. Bonettini, M. Prato, and S. Rebegoldi, *A block coordinate variable metric linesearch based proximal gradient method*, Computational Optimization and Applications, 71 (2018), pp. 5–52.

[30] S. Bonettini, M. Prato, and S. Rebegoldi, *A nested primal–dual FISTA-like scheme for composite convex optimization problems*, 2021. Preprint.

[31] S. Bonettini, M. Prato, and S. Rebegoldi, *New convergence results for the inexact variable metric forward–backward method*, Applied Mathematics and Computation, 392 (2021), pp. Paper No. 125719,21.

[32] S. Bonettini, S. Rebegoldi, and V. Ruggiero, *Inertial variable metric techniques for the inexact forward–backward algorithm*, SIAM Journal on Scientific Computing, 40 (2018), pp. A3180–A3210.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–122.

[34] R. I. Boţ and E. R. Csetnek, *An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems*, Journal of Optimization Theory and Applications, 171 (2016), pp. 600–616.

[35] R. I. Boţ, E. R. Csetnek, and S. C. László, *An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions*, EURO Journal on Computational Optimization, 4 (2016), pp. 3–25.

[36] R. H. Byrd, G. M. Chin, J. Nocedal, and F. Oztoprak, *A family of second-order methods for convex $\ell_1$-regularized optimization*, Mathematical Programming, 159 (2016), pp. 435–467.

[37] R. H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM Journal on Numerical Analysis, 26 (1989), pp. 727–739.

[38] R. H. Byrd, J. Nocedal, and F. Oztoprak, *An inexact successive quadratic approximation method for $\ell_1$ regularized optimization*, Mathematical Programming, 157 (2016), pp. 375–396.

[39] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representations of quasi-Newton matrices and their use in limited memory methods*, Mathematical Programming, 63 (1994), pp. 129–156.

[40] C. Cartis, N. I. Gould, and P. L. Toint, *On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming*, SIAM Journal on Optimization, 21 (2011), pp. 1721–1739.

[41] 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*, Advances in Computational Mathematics, 42 (2016), pp. 1287–1310.

[42] J. Chen and I. Loris, *On starting and stopping criteria for nested primal-dual iterations*, Numerical Algorithms, 82 (2019), pp. 605–621.

[43] P. Chen, J. Huang, and X. Zhang, *A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration*, Inverse Problems, 29 (2013), pp. 025011, 33.

[44] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM Journal on Numerical Analysis, 38 (2000), pp. 1200–1216.

[45] Z. Chen, A. Milzarek, and Z. Wen, *A trust-region method for nonsmooth nonconvex optimization*, arXiv preprint arXiv:2002.08513, (2020).

[46] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, *Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function*, Journal of Optimization Theory and Applications, 162 (2014), pp. 107–132.

[47] F. Clarke, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1990.

[48] C. Clason and T. Valkonen, *Introduction to nonsmooth analysis and optimization*, 2020, `https://arxiv.org/abs/2001.00216`.

[49] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-point algorithms for inverse problems in science and engineering, Springer, 2011, pp. 185–212.

[50] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Modeling & Simulation, 4 (2005), pp. 1168–1200.

[51] A. Conn, N. Gould, and P. Toint, *Trust Region Methods*, MPS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 2000.

[52] J. C. De Los Reyes, E. Loayza, and P. Merino, *Second-order orthant-based methods with enriched Hessian information for sparse $\ell_1$-optimization*, Computational Optimization and Applications, 67 (2017), pp. 225–258.

[53] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Mathematical Programming, 75 (1996), pp. 407–439.

[54] T. Deleu and Y. Bengio, *Structured sparsity inducing adaptive optimizers for deep learning*, arXiv preprint arXiv:2102.03869, (2021).

[55] J. E. Dennis and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM review, 19 (1977), pp. 46–89.

[56] J. E. Dennis and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Mathematics of Computation, 28 (1974), pp. 549–560.

[57] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Mathematical Programming, 91 (2002), pp. 201–213.

[58] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Mathematics of Operations Research, 43 (2018), pp. 919–948.

[59] E. Esser, X. Zhang, and T. F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 1015–1046.

[60] W. Fenchel, *Convex Cones, Sets, and Functions*, Princeton University, Department of Mathematics, Logistics Research Project, 1953.

[61] R. Fletcher, *A model algorithm for composite nondifferentiable optimization problems*, in Nondifferential and Variational Techniques in Optimization, Springer, 1982, pp. 67–76.

[62] K. Fountoulakis and R. Tappenden, *Robust block coordinate descent*, arXiv preprint arXiv:1407.7573, (2014).

[63] K. Fountoulakis and R. Tappenden, *A flexible coordinate descent method*, Computational Optimization and Applications, 70 (2018), pp. 351–394.

[64] P. Frankel, G. Garrigos, and J. Peypouquet, *Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates*, Journal of Optimization Theory and Applications, 165 (2015), pp. 874–900.

[65] M. P. Friedlander and G. Goh, *Efficient evaluation of scaled proximal operators*, Electronic Transactions on Numerical Analysis, 46 (2017), pp. 1–22.

[66] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain non-convex minimization problems*, International Journal of Systems Science, 12 (1981), pp. 989–1000.

[67] M. J. Gangeh, A. K. Farahat, A. Ghodsi, and M. S. Kamel, *Supervised dictionary learning and sparse representation – a review*, arXiv preprint arXiv:1502.05928, (2015).

[68] H. Ghanbari and K. Scheinberg, *Proximal quasi-Newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates*, Computational Optimization and Applications, 69 (2018), pp. 597–627.

[69] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 323–343.

[70] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, *A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems.*, in ICML (2), 2013, pp. 37–45.

[71] R. Griesse and D. A. Lorenz, *A semismooth Newton method for Tikhonov functionals with sparsity constraints*, Inverse Problems, 24 (2008), pp. 035007, 19.

[72] B. Gu, D. Wang, Z. Huo, and H. Huang, *Inexact proximal gradient methods for non-convex and non-smooth optimization*, in 32. AAAI Conference on Artificial Intelligence, 2018, pp. 3093–3100.

[73] R. Gu and A. Dogandžić, *Projected Nesterov's proximal-gradient algorithm for sparse signal recovery*, IEEE Transactions on Signal Processing, 65 (2017), pp. 3510–3525.

[74] E. T. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for $\ell_1$-minimization: Methodology and convergence*, SIAM Journal on Optimization, 19 (2008), pp. 1107–1130.

[75] J. Heinonen, *Lectures on Lipschitz analysis*, no. 100, University of Jyväskylä, 2005.

[76] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2010.

[77] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I: Fundamentals*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2013.

[78] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, *Sparse inverse covariance matrix estimation using quadratic approximation*, arXiv preprint arXiv:1306.3212, (2013).

[79] Y. Huang and H. Liu, *A Barzilai-Borwein type method for minimizing composite functions*, Numerical Algorithms, 69 (2015), pp. 819–838.

[80] K. Jiang, D. Sun, and K.-C. Toh, *An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP*, SIAM Journal on Optimization, 22 (2012), pp. 1042–1064.

[81] C. Kanzow and T. Lechner, *Efficient regularized proximal quasi-Newton methods for large-scale nonconvex composite optimization problems.* Preprint, 2022.

[82] C. Kanzow and T. Lechner, *Globalized inexact proximal Newton-type methods for nonconvex composite functions*, Computational Optimization and Applications, 78 (2021), pp. 377–410.

[83] C. Kanzow and P. Mehlitz, *Convergence properties of monotone and nonmonotone proximal gradient methods revisited*, 2021, `https://arxiv.org/abs/2112.01798`.

[84] S. Karimi and S. Vavasis, *IMRO: A proximal quasi-Newton method for solving $\ell_1$-regularized least squares problems*, SIAM Journal on Optimization, 27 (2017), pp. 583–615.

[85] N. Keskar, J. Nocedal, F. Öztoprak, and A. Waechter, *A second-order method for convex $\ell_1$-regularized optimization with active-set prediction*, Optimization Methods and Software, 31 (2016), pp. 605–621.

[86] D. Kim, S. Sra, and I. S. Dhillon, *A scalable trust-region algorithm with application to mixed-norm regression*, in ICML, 2010.

[87] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, *An interior-point method for large-scale $\ell_1$-regularized least squares*, IEEE Journal of Selected Topics in Signal Processing, 1 (2007), pp. 606–617.

[88] K. Koh, S.-J. Kim, and S. Boyd, *An interior-point method for large-scale $l_1$-regularized logistic regression*, Journal of Machine Learning Research, 8 (2007), pp. 1519–1555.

[89] N. Komodakis and J.-C. Pesquet, *Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems*, IEEE Signal Processing Magazine, 32 (2015), pp. 31–54.

[90] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, in Annales de l'Institut Fourier, vol. 48, 1998, pp. 769–783.

[91] C.-p. Lee, C. H. Lim, and S. J. Wright, *A distributed quasi-Newton algorithm for empirical risk minimization with nonsmooth regularization*, arXiv preprint arXiv:1803.01370, (2018).

[92] C.-p. Lee and S. J. Wright, *Inexact successive quadratic approximation for regularized optimization*, Computational Optimization and Applications, 72 (2019), pp. 641–674.

[93] C.-p. Lee and S. J. Wright, *Inexact variable metric stochastic block-coordinate descent for regularized optimization*, Journal of Optimization Theory and Applications, 185 (2020), pp. 151–187.

[94] D. D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401 (1999), pp. 788–791.

[95] J. D. Lee, Y. Sun, and M. A. Saunders, *Proximal Newton-type methods for minimizing composite functions*, SIAM Journal on Optimization, 24 (2014), pp. 1420–1443.

[96] D.-H. Li, M. Fukushima, L. Qi, and N. Yamashita, *Regularized Newton methods for convex minimization problems with singular solutions*, Computational Optimization and Applications, 28 (2004), pp. 131–147.

[97] G. Li and T. K. Pong, *Global convergence of splitting methods for nonconvex composite optimization*, SIAM Journal on Optimization, 25 (2015), pp. 2434–2460.

[98] G. Li and T. K. Pong, *Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods*, Foundations of Computational Mathematics, 18 (2018), pp. 1199–1232.

[99] H. Li and Z. Lin, *Accelerated proximal gradient methods for nonconvex programming*, in Advances in Neural Information Processing Systems, 2015, pp. 379–387.

[100] J. Li, M. S. Andersen, and L. Vandenberghe, *Inexact proximal Newton methods for self-concordant functions*, Mathematical Methods of Operations Research, (2016), pp. 1–23.

[101] X. Li, D. Sun, and K.-C. Toh, *A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems*, SIAM Journal on Optimization, 28 (2018), pp. 433–458.

[102] C.-J. Lin, *Projected gradient methods for nonnegative matrix factorization*, Neural Computation, 19 (2007), pp. 2756–2779.

[103] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Mathematical Programming, 45 (1989), pp. 503–528.

[104] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les équations aux dérivées partielles, 117 (1963), pp. 87–89.

[105] Z.-Q. Luo and P. Tseng, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM Journal on Optimization, 2 (1992), pp. 43–54.

[106] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization, 30 (1992), pp. 408–425.

[107] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: a general approach*, Annals of Operations Research, 46 (1993), pp. 157–178.

[108] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online dictionary learning for sparse coding*, in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 689–696.

[109] L. Meier, S. Van De Geer, and P. Bühlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70 (2008), pp. 53–71.

[110] C. A. Micchelli, L. Shen, and Y. Xu, *Proximity algorithms for image models: denoising*, Inverse Problems, 27 (2011), pp. 045009,30.

[111] A. Milzarek, *Numerical methods and second order theory for nonsmooth problems*, PhD thesis, Technische Universität München, 2016.

[112] A. Milzarek and M. Ulbrich, *A semismooth Newton method with multidimensional filter globalization for $l_1$-optimization*, SIAM Journal on Optimization, 24 (2014), pp. 298–333.

[113] H. Mine and M. Fukushima, *A minimization method for the sum of a convex function and a continuously differentiable function*, Journal of Optimization Theory and Applications, 33 (1981), pp. 9–23.

[114] B. S. Mordukhovich, X. Yuan, S. Zeng, and J. Zhang, *A globally convergent proximal Newton-type method in nonsmooth convex optimization*, arXiv preprint arXiv:2011.08166, (2020).

[115] J. J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien*, Comptes rendus hebdomadaires des séances de l'Académie des sciences, 255 (1962), pp. 2897–2899.

[116] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société mathématique de France, 93 (1965), p. 273–299.

[117] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM Journal on Scientific and Statistical Computing, 4 (1983), pp. 553–572.

[118] P. Q. Muoi, D. N. Hào, P. Maass, and M. Pidcock, *Semismooth Newton and quasi-Newton methods in weighted $\ell_1$-regularization*, Journal of Inverse and Ill-Posed Problems, 21 (2013), pp. 665–693.

[119] S. Nakayama, Y. Narushima, and H. Yabe, *Inexact proximal memoryless quasi-Newton methods based on the Broyden family for minimizing composite functions*, Computational Optimization and Applications, 79 (2021), pp. 127–154.

[120] Y. Nesterov, *Gradient methods for minimizing composite objective function*, CORE Discussion paper #2007/76, CORE, 2007.

[121] Y. Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming, 140 (2013), pp. 125–161.

[122] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Mathematics of Computation, 35 (1980), pp. 773–782.

[123] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, 2006.

[124] G. Nolet, *A breviary of seismic tomography*, A Breviary of Seismic Tomography, (2008).

[125] J. Norbury, *Big Panda and Tiny Dragon*, Mandala Publishing, 2021.

[126] P. Ochs, *Local convergence of the heavy-ball method and iPiano for non-convex optimization*, arXiv e-prints arXiv:1606.09070, (2016).

[127] P. Ochs, T. Brox, and T. Pock, *iPiasco: Inertial proximal algorithm for strongly convex optimization*, Journal of Mathematical Imaging and Vision, 53 (2015), pp. 171–181.

[128] P. Ochs, Y. Chen, T. Brox, and T. Pock, *iPiano: Inertial proximal algorithm for nonconvex optimization*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1388–1419.

[129] P. Ochs and T. Pock, *Adaptive FISTA for nonconvex optimization*, SIAM Journal on Optimization, 29 (2019), pp. 2482–2503.

[130] P. A. Olsen, F. Öztoprak, J. Nocedal, and S. J. Rennie, *Newton-like methods for sparse inverse covariance estimation.*, in NIPS, vol. 25, Citeseer, 2012, pp. 764–772.

[131] W. Ouyang and A. Milzarek, *A trust region-type normal map-based semismooth Newton method for nonsmooth nonconvex composite optimization*, arXiv preprint arXiv:2106.09340, (2021).

[132] N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in Optimization, 1 (2014), pp. 127–239.

[133] P. Patrinos and A. Bemporad, *Proximal Newton methods for convex composite optimization*, in 52nd IEEE Conference on Decision and Control, IEEE, 2013, pp. 2358–2363.

[134] P. Patrinos, L. Stella, and A. Bemporad, *Forward-backward truncated Newton methods for convex composite optimization*, arXiv preprint arXiv:1402.6655, (2014).

[135] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Mathematics of Operations Research, 18 (1993), pp. 227–244.

[136] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Mathematical Programming, 58 (1993), pp. 353–367.

[137] L. Qi and J. Sun, *A trust region algorithm for minimization of locally Lipschitzian functions*, Mathematical Programming, 66 (1994), pp. 25–43.

[138] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, *High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence*, Electronic Journal of Statistics, 5 (2011), pp. 935–980.

[139] S. Rebegoldi, S. Bonettini, and M. Prato, *A Bregman inexact linesearch–based forward–backward algorithm for nonsmooth nonconvex optimization*, in Journal of Physics: Conference Series, vol. 1131, IOP Publishing, 2018, p. 012013.

[140] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[141] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 2015.

[142] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317, Springer Science & Business Media, 2009.

[143] K. Scheinberg, D. Goldfarb, and X. Bai, *Fast first-order methods for composite convex optimization with backtracking*, Foundations of Computational Mathematics, 14 (2014), pp. 389–417.

[144] K. Scheinberg and I. Rish, *SINCO – A greedy coordinate ascent method for sparse inverse covariance selection problem*, Preprint, (2009).

[145] K. Scheinberg and X. Tang, *Practical inexact proximal quasi-Newton method with global complexity analysis*, Mathematical Programming, 160 (2016), pp. 495–529.

[146] M. Schmidt, N. L. Roux, and F. R. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Advances in Neural Information Processing Systems, 2011, pp. 1458–1466.

[147] D. Steck and C. Kanzow, *Regularization of limited memory quasi-Newton methods for large-scale nonconvex minimization*, arXiv preprint arXiv:1911.04584, (2019).

[148] L. Stella, A. Themelis, and P. Patrinos, *Forward–backward quasi-Newton methods for nonsmooth optimization problems*, Computational Optimization and Applications, 67 (2017), pp. 443–487.

[149] W. W. Symes, *The seismic reflection inverse problem*, Inverse Problems, 25 (2009), p. 123008.

[150] A. Themelis, L. Stella, and P. Patrinos, *Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms*, SIAM Journal on Optimization, 28 (2018), pp. 2274–2303.

[151] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological), 58 (1996), pp. 267–288.

[152] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher, *A proximal Newton framework for composite minimization: Graph learning without Cholesky decompositions and matrix inversions*, in International Conference on Machine Learning, 2013, pp. 271–279.

[153] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher, *An inexact proximal path-following algorithm for constrained convex minimization*, SIAM Journal on Optimization, 24 (2014), pp. 1718–1745.

[154] Q. Tran-Dinh, A. T. Kyrillidis, and V. Cevher, *Composite self-concordant minimization.*, Journal of Machine Learning Research, 16 (2015), pp. 371–416.

[155] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization.* `http://newton.kias.re.kr/~yswcs/ysweb/cgd.pdf`, 2006. unpublished.

[156] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 117 (2009), pp. 387–423.

[157] K. Ueda and N. Yamashita, *Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization*, Applied Mathematics and Optimization, 62 (2010), pp. 27–46.

[158] K. Ueda and N. Yamashita, *A regularized Newton method without line search for unconstrained optimization*, Computational Optimization and Applications, 59 (2014), pp. 321–351.

[159] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 2011.

[160] T. Valkonen, *A primal–dual hybrid gradient method for nonlinear operators with applications to MRI*, Inverse Problems, 30 (2014), pp. 055012, 45.

[161] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, *Accelerated and inexact forward-backward algorithms*, SIAM Journal on Optimization, 23 (2013), pp. 1607–1633.

[162] J. Walker, *A Primer on Wavelets and Their Scientific Applications*, Studies in Advanced Mathematics, CRC Press, 2008.

[163] S. J. Wright, *Optimization algorithms for data analysis*, The Mathematics of Data, 25 (2018), pp. 49–97.

[164] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Transactions on Signal Processing, 57 (2009), pp. 2479–2493.

[165] L. Yang, *Proximal gradient method with extrapolation and line search for a class of nonconvex and nonsmooth problems*, arXiv preprint arXiv:1711.06831, (2017).

[166] G.-X. Yuan, C.-H. Ho, and C.-J. Lin, *An improved GLMNET for l1-regularized logistic regression*, Journal of Machine Learning Research, 13 (2012), pp. 1999–2030.

[167] M.-C. Yue, Z. Zhou, and A. M.-C. So, *A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property*, Mathematical Programming, 174 (2019), p. 327–358.

[168] H. Zhang, J. Jiang, and Z.-Q. Luo, *On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems*, Journal of the Operations Research Society of China, 1 (2013), pp. 163–186.

[169] K. Zhong, I. E.-H. Yen, I. S. Dhillon, and P. K. Ravikumar, *Proximal quasi-Newton for computationally intensive $\ell_1$-regularized M-estimators*, in Advances in Neural Information Processing Systems, 2014, pp. 2375–2383.

[170] Z. Zhou and A. M.-C. So, *A unified approach to error bounds for structured convex optimization problems*, Mathematical Programming, 165 (2017), pp. 689–728.

[171] M. Zulfiquar Ali Bhotto, M. O. Ahmad, and M. Swamy, *An improved fast iterative shrinkage thresholding algorithm for image deblurring*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 1640–1657.