Geometric numerical integration for optimisation

Erlend Skaldehaug Riis

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Queens' College September 2019

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

Erlend Skaldehaug Riis
September 2019

Abstract

In this thesis, we study geometric numerical integration for the optimisation of various classes of functionals. Numerical integration and the study of systems of differential equations have received increased attention within the optimisation community in the last decade, as a means for devising new optimisation schemes as well as to improve our understanding of the dynamics of existing schemes. Discrete gradient methods from geometric numerical integration preserve structures of first-order gradient systems, including the dissipative structure of schemes such as gradient flows, and thus yield iterative methods that are unconditionally dissipative, i.e. decrease the objective function value for all time steps.

We look at discrete gradient methods for optimisation in several settings. First, we provide a comprehensive study of discrete gradient methods for optimisation of continuously differentiable functions. In particular, we prove properties such as well-posedness of the discrete gradient update equation, convergence rates, convergence of the iterates, and propose methods for solving the discrete gradient update equation with superior stability and convergence rates. Furthermore, we present results from numerical experiments which support the theory.

Second, motivated by the existence of derivative-free discrete gradients, and seeking to solve nonsmooth optimisation problems and more generally black-box problems, including for parameter optimisation problems, we propose methods based on the Itoh–Abe discrete gradient method for solving nonconvex, nonsmooth optimisation problems with derivative-free methods. In this setting, we prove well-posedness of the method, and convergence guarantees within the nonsmooth, nonconvex Clarke subdifferential framework for locally Lipschitz continuous functions. The analysis is shown to hold in various settings, namely in the unconstrained and constrained setting, including epi-Lipschitzian constraints, and for stochastic and deterministic optimisation methods.

Building on the work of derivative-free discrete gradient methods and the concept of structure preservation in geometric numerical integration, we consider discrete gradient methods applied to other differential systems with dissipative structures. In particular, we study the inverse scale space flow, linked to the well-known Bregman methods, which are central to variational optimisation problems and regularisation methods for inverse problems.

In this setting, we propose and implement derivative-free schemes that exploit structures such as sparsity to achieve superior convergence rates in numerical experiments, and prove convergence guarantees for these methods in the nonsmooth, nonconvex setting. Furthermore, these schemes can be seen as generalisations of the Gauss-Seidel method and successive-over-relaxation.

Finally, we return to parameter optimisation problems, namely nonsmooth bilevel optimisation problems, and propose a framework to employ first-order methods for these problems, when the underlying variational optimisation problem admits a nonsmooth structure in the partial smoothness framework. In this setting, we prove piecewise differentiability of the parameter-dependent solution mapping, and study algorithmic differentiation approaches to evaluating the derivatives. Furthermore, we prove that the algorithmic derivatives converge to the implicit derivatives. Thus we demonstrate that, although some parameter tuning problems must inevitably be treated as black-box optimisation problems, for a large number of variational problems one can exploit the structure of nonsmoothness to perform gradient-based bilevel optimisation.

This thesis is dedicated to my parents.

Acknowledgements

After four years in Cambridge that have been beyond wonderful, and at the end of an occasionally chaotic PhD journey, there are many people I would like to thank for being in my life during this period, both in personal and professional capacities.

First of all, I am deeply grateful and indebted to my supervisor Carola Schönlieb. The thesis would not have happened without your academic guidance, support, generosity, and patience, and I count myself as exceptionally lucky to have had you as my doctoral supervisor, and to be a part of the Cambridge Image Analysis group.

I would sincerely like to thank everyone in the CIA by name, but fear at this point that the names alone would fill a page . However, I am very glad to have been in a research group in which academic and social isolation have been such alien concepts, and from which I have many fond memories. Matthias Ehrhardt, thank you for the enjoyable collaborations and discussions, as well as for encouraging me to apply for the LMS Early Career fellowship. Martin Benning, thank you for our collaboration, your guidance on how to 'Bregmanise' discrete gradient methods, and your positive energy. Tamara Großmann, thank you for our collaboration—amidst the ceaseless debugging of MATLAB toolboxes, it was great to work with you. Joana Grah, thank you for all the BBQs you hosted, and the lunches at Robinson. Finally, thank you Jingwei Liang for taking the time to talk through various aspects of partial smoothness with me.

My greatest thanks to Reinout Quispel and Robert McLachlan for the wonderful hospitality they showed me during my trips to Melbourne and Palmerston North. I am grateful for being introduced to geometric numerical integration during my stays with you. Furthermore, thank you to Reinout Quispel and Torbjørn Ringholm for our collaborations and discussions. Thanks as well to everyone in my cohort of the CCA, and an extra thanks to Eardi Lila for your former, involuntary role as my go-to MATLAB solver. I am grateful to Anders Hansen, for his time and guidance as my supervisor during the first year of CCA, and to Sarah Bohndiek and James Joseph for all the help and discussions during the project on photoacoustic tomography. Thank you to Andrew Stuart for all his support during my master's studies, ultimately fooling CCA into accepting my application. Thank you to the CCA

directors for accepting me onto the programme. Finally, I am thankful to Cantab Capital Institute for the Mathematics of Information for funding the latter part of my PhD.

I sincerely thank the thesis examiners Prof. Elena Celledoni and Dr. Hamza Fawzi for carefully reading the manuscript and for the invaluable feedback and discussions during my viva.

Andria Pancrazi and Jenny Leivadarou, thank you for all the wonderful memories from my first year in Cambridge, and an extra shout-out to Jenny for her valiant efforts in convincing me to stick with the PhD. Thank you to Joel Mandella for visiting Cambridge twice, and of course the rest of the Mandellas, for all the times we have had together during the past 11 years. Thank you to my good friends Njål, Nicolai, and Grete for making the trip from Oslo.

Last but not least, I would have not have completed this programme, let alone get into Cambridge, without the love and support of my parents through all the years, even in spite of all those slanderous stories I wrote in school about flawed parents. You make excellent parents, and promise I will take up Norwegian again some day. My sister Oda and brother Pål, at the risk of jinxing the end of my PhD, I look forward to demanding you call me 'Dr Erlend'. My partner Freya, largely thanks to you, I have stayed relatively sane during the final weeks. Throughout my PhD, your ambition and academic abilities have inspired me to strive to work harder, which this thesis hopefully reflects.

Table of contents

List of figures

List of tables

Nomenclature

Differentials

$C^k(\mathbb{R}^n;\mathbb{R}^m)$	The set of k times continuously differentiable functions $f:\mathbb{R}^n\to\mathbb{R}^m$.
$C^k_c(\mathbb{R}^n;\mathbb{R}^m)$	The set of functions in $C^k(\mathbb{R}^n;\mathbb{R}^m)$ with compact support.
∇f	Gradient of $f:\mathbb{R}^n\to\mathbb{R}$.
Df	The first-order differential of f.
$\nabla_x f(x,\vartheta)$	Gradient with respect to spatial vector x.
$D_\vartheta f(x,\vartheta)$	Differential with respect to parameters ϑ.
$\nabla^2 f$	Hessian of $f:\mathbb{R}^n\to\mathbb{R}$, i.e. $\nabla^2 f = \left(\dfrac{\partial^2 f}{\partial x_i \partial x_j}\,\dfrac{\partial^2 f}{\partial x_i x_j^2}\right)_{i,j=1,\dots,n}$.
$f'(\cdot;d)$	The directional derivative of f in the direction d.
∇x	Discretised spatial gradient of the vector x.
$\dot{x}(t)$	The temporal derivative of a vector-valued function x.
$\mathcal{E}$	Symmetrised gradient.

Functions

$\Gamma_0(\mathbb{R}^n)$	The set of convex, proper, lower semicontinuous functions on $\mathbb{R}^n$.
$\operatorname{gph} f$	The graph of f.
$\operatorname{dom} f$	The effective domain of f.
$\operatorname{epi} f$	The epigraph $\{(x,\alpha)\ :\ \alpha \geq f(x)\}$ of f.
$\operatorname{supp} f$	The support of f.

δ_Ω	The indicator function of $\Omega \subset \mathbb{R}^n$.
$\operatorname{sgn}(x)$	The sign function.
$f \circ g$	The composition of functions f and g.
f^*	The convex conjugate of f.
$J \overset{+}{\vee} G$	The infimal convolution of J and G.

Generalised differentials

∂f	The (general) subdifferential of f.
$\partial^C f$	The Clarke subdifferential of f.
$\widehat{\partial} f$	The regular subdifferential of f.
$\partial^\infty f$	The horizon subdifferential of f.
$f^o(\cdot;d)$	The Clarke directional derivative of f in the direction d.
D_f, D_f^{symm}	The Bregman distance and symmetric Bregman distance.
$\overline{\nabla} f$	A discrete gradient of f.
$\operatorname{SGN}(x)$	The (set-valued) subdifferential of $\lvert x\rvert$.

Linear algebra

A^*	The adjoint of a matrix A.
A^H	Hermitian adjoint of matrix A.
I_n	The $n \times n$ identity matrix.
0_n	The $n \times n$ zero matrix.
$\operatorname{rank}(A)$	Rank of matrix A.
$\sigma(A)$	Spectrum of matrix A.
$\rho(A)$	Spectral radius of matrix A.
κ_A	The condition number of matrix A.
$A^\dagger$	The Moore-Penrose pseudoinverse of matrix A.

$\ker A$	The kernel of a matrix A.

Miscellaneous

e^i	The ith coordinate vector of $\mathbb{R}^n$.
$\int f \, \mathrm{d}s$	The integral of f with respect to s.
$b(\mod a)$	The modulo operation.
$\mathbb{E}_\xi$	The expectation with respect to distribution ξ.
$f(x) = o(g(x))$	Little o: $\lim_{\|x\|\to 0} \|f(x)\|/\|g(x)\| = 0$.
$f(x) = O(g(x))$	Big O: $\limsup_{\|x\|\to 0} \|f(x)\|/\|g(x)\| \leq C < \infty$.

Numerical sets

$\mathbb{N}$	The set of natural numbers.
$\mathbb{Z}$	The set of integers.
$\mathbb{R}$	The set of real numbers.
$\overline{\mathbb{R}}$	The extended real numbers $\mathbb{R} \cup \{\pm\infty\}$.
$\mathbb{R}_{\geq 0}$	The set of nonnegative, real numbers.
$\mathbb{C}$	The set of complex numbers.
$\mathbb{Q}$	The set of rational numbers.
$\mathbb{R}^{m,n}$	The set of $m \times n$ real matrices.
$\mathbb{C}^{m,n}$	The set of complex-valued $m \times n$ matrices.
$B_\varepsilon(x)$	The open ball of radius $\varepsilon > 0$ at x.
$\overline{B}_\varepsilon(x)$	The closed ball of radius $\varepsilon > 0$ around x.
S^{n-1}	The unit sphere in $\mathbb{R}^n$.

Norms

$\|\cdot\|$, $\langle \cdot, \cdot \rangle$	The Euclidean norm and associated inner product on $\mathbb{R}^n$.
$\|\cdot\|_p$	The ℓ^p-norm for $p \in [0, +\infty]$.

$\|\cdot\|_F$	The Frobenius norm for matrices.
$\lvert x\rvert$	The absolute value of x.
$\|x\|_{1,2}$	The group lasso norm $\sum_{i=1}^n \|x^i\|$.
TV	The total variation seminorm $\|\nabla\cdot\|_{1,2}$.
TGV	The total generalised variation seminorm.
$\lvert x\rvert_\gamma$	The smoothed absolute value $\sqrt{x^2+\gamma}$.
$\|x\|_\gamma$	The smoothed norm $\sqrt{\|x\|^2+\gamma}$.
$\|x\|_{1,2,\gamma}$	The smoothed group lasso norm $\sum_{i=1}^n \|x^i\|_\gamma$.
TV_γ	The smoothed total variation seminorm $\|\nabla\cdot\|_{1,2,\gamma}$.

Riemannian manifolds

$T_x\mathcal{M}$	The tangent space of manifold $\mathcal{M}$.
$N_x\mathcal{M}$	The normal space of manifold $\mathcal{M}$.
$\nabla_{\mathcal{M}} f$	The Riemannian gradient of f along $\mathcal{M}$.
$\nabla^2_{\mathcal{M}} f$	The Riemannian Hessian of f along $\mathcal{M}$.
$\mathfrak{W}_x$	The Weingarten map of $\mathcal{M}$ at x.

Set operations

X^n	The n-Cartesian product of a set X.
$U \Subset V$	U is compactly contained in V.
$X^\perp$	The orthogonal complement of a set $X \subset \mathbb{R}^n$.
$[\mathbf{x},\mathbf{y}]$	The line segment $\{\lambda\mathbf{x}+(1-\lambda)\mathbf{y} \,:\, \lambda \in [0,1]\}$.
$\mathrm{cl}\,\Omega$	The closure of $\Omega \subset \mathbb{R}^n$.
$\mathrm{int}\,\Omega$	The interior of $\Omega \subset \mathbb{R}^n$.
$\mathrm{bd}\,\Omega$	The topological boundary of Ω, $\mathrm{cl}\,\Omega \setminus \mathrm{int}\,\Omega$.
$\mathrm{diam}\,\Omega$	The diameter of $\Omega \subset \mathbb{R}^n$.

$\mathrm{dist}(\Omega, x)$	The Euclidean distance from $x \in \mathbb{R}^n$ to $\Omega \subset \mathbb{R}^n$.
N_x	Neighbourhood of $x \in \mathbb{R}^n$.
$\mathrm{ri}\,\Omega$	The relative interior of $\Omega \subset \mathbb{R}^n$.
$\mathrm{rbd}\,\Omega$	The relative topological boundary of Ω, $\mathrm{cl}\,\Omega \setminus \mathrm{ri}\,\Omega$.
$\mathrm{co}\,\Omega$	The convex hull of $\Omega \subset \mathbb{R}^n$.
$\mathrm{aff}\,\Omega$	The affine hull of $\Omega \subset \mathbb{R}^n$.
$\mathrm{par}\,\Omega$	The subspace parallel to $\mathrm{aff}\,\Omega$.
P_X	The projection operator of a closed, convex set $X \subset \mathbb{R}^n$.
T_Ω^C	The Clarke tangent cone of $\Omega \subset \mathbb{R}^n$.
T_Ω^H	The hypertangent cone of $\Omega \subset \mathbb{R}^n$.
Ω^∞	The horizon cone of Ω.
$\widehat{N}_\Omega$	The regular normal space of Ω.
N_Ω	The (general) normal space of Ω.

Abbreviations

ADMM	Alternating direction method of multipliers
BSOR	Bregman successive-over-relaxation
CCD	Cyclic coordinate descent
CIA	Cyclic Itoh–Abe method
DG	Discrete gradient
FISTA	Fast iterative shrinkage-thresholding algorithm
ISS	Inverse scale space
KŁ	Kurdyka–Łojasiewicz
MADS	Mesh adaptive direct search
MRF	Markov random field

MRI	Magnetic resonance imaging
ND	Nondegeneracy
ODE	Ordinary differential equation
PDE	Partial differential equation
PDHG	Primal-dual hybrid gradient
PŁ	Polyak–Łojasiewicz
RCD	Randomised coordinate descent
RIA	Randomised Itoh–Abe method
SOR	Successive-over-relaxation
SSIM	Structural similarity index

Chapter 1

Introduction

In this thesis, we consider methods for solving various classes of optimisation problems, by combining tools from geometric numerical integration and nonsmooth optimality analysis. In this opening chapter, we provide the context and rationale for the research, and outline the contributions.

At the core of mathematical optimisation is the problem

$$\min_{x \in \mathcal{X}} F(x), \quad \text{i.e. find } x \text{ that minimises } F : \mathcal{X} \to \mathbb{R}. \tag{1.1}$$

This objective is notably simple. However, solving (1.1) can be arbitrarily difficult depending on the properties of the objective function F, and whether or not such a problem is solvable will have implications for the feasibility of various scientific applications. Because of this, optimisation theory is a broad and dynamic field whose developments both influence and are influenced by advances in science and engineering.

In contemporary optimisation problems, one often encounters challenges such as high dimensionality of $\mathcal{X}$, nonsmoothness of F, and nonlinearities in $\mathcal{X}$ and the gradient ∇F. In part, this is due to the rise of big data and machine learning, as well as sophisticated sparsity and regularisation models in signal processing which invoke nonsmooth functions to promote efficient signal representations. This has lead to a surge of interest in first-order optimisation methods, which scale better than higher-order methods with respect to the dimension of $\mathcal{X}$, and which combine naturally with nonsmooth optimisation via proximal methods. While this has led to rapid developments of mathematical optimisation theory in recent decades, there is as much as ever a demand for the development of algorithmic frameworks in optimisation for handling complex, nonlinear dynamics in an efficient manner.

As is often the case in mathematics, a promising source of ideas and developments in optimisation is to start with perspectives and techniques from other areas of mathematics.

One such area which has played an integral role to optimisation in recent years is numerical integration. This is the study of numerical methods for solving systems of differential equations. The subfield geometric numerical integration is the systematic study of geometric structures of differential systems and structure-preserving numerical methods. A recurring theme of this thesis is how techniques from geometric numerical integration can be used to solve a wide range of optimisation problems, through the preservation of energy dissipation.

For the remainder of this chapter, we will provide brief introductions to first-order optimisation methods, variational optimisation problems from signal processing which motivate our research, and numerical integration applied to optimisation methods, and finally give an outline and summarise the contributions of this thesis in Sections 1.2 & 1.3

1.1 An overview of optimisation and numerical integration

1.1.1 First-order optimisation methods

The two most important building blocks to first-order optimisation methods are *explicit* and *implicit gradient descent*. For a differentiable function $F : \mathbb{R}^n \to \mathbb{R}$, starting point $x^0 \in \mathbb{R}^n$, and strictly positive time steps $(\tau_k)_{k\in\mathbb{N}}$, (explicit) gradient descent is given by

$$x^{k+1} = x^k - \tau_k \nabla F(x^k), \tag{1.2}$$

while implicit gradient descent is given by

$$x^{k+1} = x^k - \tau_k \nabla F(x^{k+1}). \tag{1.3}$$

As the names suggest, the former update is explicit, while the latter is implicit, i.e. one must solve (1.3) with respect to x^{k+1}. Note that this is equivalent to

$$0 = \nabla\left(F(x^{k+1}) + \frac{1}{2\tau_k}\|x^k - x^{k+1}\|^2\right).$$

It follows that if $y \mapsto F(y) + \|y - x^k\|^2/(2\tau_k)$ is convex, then x^{k+1} solves (1.3) if and only if it solves

$$x^{k+1} = \underset{y\in\mathbb{R}^n}{\arg\min} F(y) + \frac{1}{2\tau_k}\|y - x^k\|^2. \tag{1.4}$$

This mapping is called the *proximal mapping* of F at x^k. Observe that this latter formulation gives us a notion of implicit gradient descent updates for nondifferentiable functions, provided the minimisation problem in (1.4) is computationally tractable. Of course, by itself this observation is not helpful, as solving (1.4) will in general be of similar difficulty as solving the problem (1.1), which in the end is what we are interested in. However, the formulation (1.4) is crucial for several reasons, some of which we will discuss now, and some which will be touched upon at different stages of the thesis.

We first emphasise that for nonsmooth variational optimisation problems, one is often able to *split* the objective function $F : \mathbb{R}^n \to \mathbb{R}$ into different terms, each of which is either continuously differentiable or admits computationally tractable updates for (1.4). From hereon, we refer to functions in the latter category as *simple*. A well-known example of a simple function is the ℓ^1-norm $\|x\|_1 := \sum_{i=1}^n |x_i|$, for which the update (1.4) corresponds to the *soft shrinkage* operator

$$S(x^k, \tau_k) := \operatorname{sgn}(x^k) \max\{|x^k| - \tau_k, 0\}, \tag{1.5}$$

where sgn is the sign operator,

$$\operatorname{sgn}(x) := \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0, \end{cases}$$

and sgn, max and their products are evaluated elementwise. Another example is that of indicator functions of convex, nonempty, closed sets $C \subset \mathbb{R}^n$,

$$\delta_C(x) := \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{otherwise}, \end{cases} \tag{1.6}$$

for which (1.4) corresponds to the *projection mapping*

$$P_C(x^k) := \arg\min_{y \in C} \frac{1}{2} \|y - x^k\|^2.$$

Second, as can be inferred directly from the equation, the update (1.4) is *unconditionally dissipative* with respect to the time step τ_k in the sense that $F(x^{k+1}) \leq F(x^k)$ for any $\tau_k > 0$. This is in contrast to explicit gradient descent, in which the time step must be restricted according to the first-order regularity of F in order to ensure energy decrease.

We return to proximal mappings in Section 2.4. For further details on proximal mappings, we refer to a survey by Combettes & Pesquet [55] and a monograph by Parikh & Boyd [172].

We point out that implicit gradient descent is also known as the *proximal point algorithm* [101, 189].

To further emphasise that the proximal mapping (1.4) exists within a powerful framework for nonsmooth optimisation, we conclude this section by pointing out that there are various first-order splitting algorithms available for solving nonsmooth variational problems, based on combinations of implicit and explicit gradient descent steps.

These include forward-backward (FB) type methods when F is the sum of a smooth function and a simple function [14, 57, 173, 85], and the Douglas–Rachford method [75, 79, 138] and the related alternating direction method of multipliers (ADMM) when F consists of two simple functions [27, 93]. Finally, primal-dual hybrid gradient (PDHG) methods [50, 226] are used for optimisation problems involving terms that are neither smooth nor simple but can be transformed into so-called saddle-point problems, and for which each term is either smooth or simple (see Section 2.4).

We return to PDHG and FB methods in Chapter 7 and Chapter 8, when we consider them for solving the lower-level problem of a bilevel optimisation problem (to be defined in the following section).

For a general treatment of convex analysis and monotone operator theory, see [12].

1.1.2 Variational optimisation problems

In what follows, we summarise common variational optimisation problems encountered in signal and image processing, and more generally inverse problems.

An inverse problem seeks to recover a signal $x \in \mathbb{R}^n$ from data $f^\delta \in \mathbb{R}^l$, via a forward model $G : \mathbb{R}^n \to \mathbb{R}^l$, i.e. such that x solves

$$f^\delta = G(x) + \delta, \tag{1.7}$$

where δ represents noise in the data f^δ.

Before we introduce variational regularisation models for solving inverse problems, we motivate their necessity with the concept of *ill-posedness*. An inverse problem is said to be *well-posed* if it admits a *unique* solution $x^* \in \mathbb{R}^n$ such that $G(x^*) = f^\delta$, and which is stable with respect to perturbations in the data f^δ. Due to the presence of noise in the data, as well as other uncertainties, such as errors in the forward model G, a direct inversion of G might lead to an inaccurate reconstruction. Furthermore, G is often ill-conditioned, meaning that even small amounts of noise in the data can lead to significant artefacts in the reconstruction. For these reasons, many inverse problems are ill-posed.

When an inverse problem is ill-posed, one needs to formulate a reconstruction model that finds a good signal match for the data, i.e. x such that $G(x)$ is close to f^δ, while simultaneously ensuring that the reconstruction adequately accounts for the uncertainty and instability in the inverse problem.

One popular approach is to use variational regularisation models, wherein one solves the variational optimisation problem

$$\arg\min_{x\in\mathbb{R}^n} F(G(x), f^\delta) + R(x, \vartheta). \tag{1.8}$$

Here $F : \mathbb{R}^l \times \mathbb{R}^l$ is a *data fidelity* term which measures the discrepancy between $G(x)$ and f^δ, one example being $F(x,y) = \|x-y\|^2/2$, and $R(\cdot,\vartheta)$ a *regularisation term*, which promotes known information about the true signal, such as smooth regions and sharp edges of images, or signal sparsity in some coordinate frame.

We mention some regularisers $R(\cdot,\vartheta)$ commonly used in image and signal processing.

The aforementioned ℓ^1-norm weighted by $\vartheta \in \mathbb{R}_{\geq 0}$, i.e. $\vartheta\|\cdot\|_1$ for $\vartheta \geq 0$, is a popular choice for promoting sparsity. We elaborate on this in the next subsection.

The *total variation* (TV) seminorm [48] in $\mathbb{R}^n$ is defined as

$$TV_\vartheta(x) := \vartheta\|\nabla x\|_{1,2}, \quad \vartheta \geq 0. \tag{1.9}$$

Here $\nabla \in \mathbb{R}^{2n,n}$ denotes the *discretised spatial gradient* for vectors in $\mathbb{R}^n$, as defined in [47], and $\|\cdot\|_{1,2}$ is the *group Lasso norm*, which for $z \in \mathbb{R}^{n,m}$ is defined as $\|z\|_{1,2} := \sum_{i=1}^n \|z^i\|$, where z^i denotes the ith row vector of z. This is the discretised version of total variation. We introduce its original, continuous formulation in Section 4.5.2. This regulariser is popular in image processing for its ability to preserve edges while penalising noise [196].

In spite of the prevalence of the TV regulariser, one of its drawbacks is that it promotes piecewise constant features, which can lead to so-called staircasing effects when the input image has linear features. To remove noise while promoting linear and higher-order features in images, Bredies et al. proposed the *total generalised variation* (TGV) seminorm [29]. In the discretised setting, this is given by

$$\mathrm{TGV}_{\vartheta_1,\vartheta_2}(x) := \min_{v\in\mathbb{R}^{2n,n}} \vartheta_1\|\nabla x\|_{1,2} + \vartheta_2\|\mathcal{E}v\|_{1,2}, \quad \vartheta_1 \geq 0, \vartheta_2 \geq 0, \tag{1.10}$$

where $\mathcal{E}$ is the symmetrised gradient [29]. Similarly as for TV, the continuous formulation is given in Section 4.5.2.

We emphasise that regularisers are often parametrised, and the reconstruction of (1.8) therefore depends on the parameter choice ϑ. A common practical and theoretical challenge in solving inverse problems is to tune these parameters appropriately. This leads to the class of *bilevel problems* which we will discuss later in this chapter, and which is a recurring topic in this thesis.

For a review of regularisation methods for inverse problems, see [18, 83].

Nonsmoothness in variational problems

A central topic in this thesis is nonsmoothness in optimisation problems. As can be observed in the previous subsection, nonsmoothness is also a common feature for variational regularisation models—e.g. the three aforementioned regularisers are nonsmooth. In what follows, we will further motivate why nonsmoothness plays an important role in signal processing, and why frameworks and methods that account for nonsmoothness in optimisation problems are worth pursuing.

We take the example of signal sparsity and compressed sensing [73]. Since the seminal works of Daubechies, Meyer, et al. on wavelets and compressibility [6, 64, 150, 151], it is well-known that real-world signals such as audio recordings and images are highly *compressible* in certain bases and dictionaries. Compressibility in this context means that in some basis, such as the wavelet basis, the signal $y \in \mathbb{R}^m$ can to a high level of accuracy be represented by a small number of bases vectors s relative to its ambient dimensionality m, i.e. $s \ll m$. Given a signal $y \in \mathbb{R}^m$ and a dictionary matrix $W \in \mathbb{R}^{m,n}$ which can be seen as a transformation from the sparsity basis to the basis of y, it is therefore of interest to find a sparse vector $x \in \mathbb{R}^n$ such that $Wx \approx y$. A reasonable variational approach to this is to solve

$$\frac{1}{2}\|Wx-y\|^2 + \vartheta\|x\|_0,$$

where $\|x\|_0 = |\operatorname{supp}(x)|$ is the number of nonzero elements in x, and $\operatorname{supp}(x) := \{i \,:\, x_i \neq 0\}$. However, as $\|\cdot\|_0$ is nonconvex and discontinuous, this problem is computationally intractable [41]. Alternatively, one could solve the optimisation problem

$$\frac{1}{2}\|Wx-y\|^2 + \vartheta\|x\|_1.$$

In addition to being convex and continuous, recall that $\|\cdot\|_1$ is simple, so this problem is amenable to proximal splitting methods.

Compressibility of signals has proven to be a powerful idea, with several applications. In *compressed sensing* [41, 73], one exploits the inherent redundancy of information in signals to allow for subsampling of f^δ, i.e. only observe a subset of the elements of f^δ, while still guaranteeing exact recovery of the ground truth signal x. Furthermore, Candès et al. show that these recovery guarantees hold if one replaces $\|\cdot\|_0$ with the convex relaxation $\|\cdot\|_1$ like we did above [41, Theorem 1.3]. This has implications e.g. for magnetic resonance imaging (MRI), as it allows for a significant reduction in scanning time [143].

Another example where data compressibility has played a crucial role is in the domains of big data, where the mere scale of the data collected renders compression a practical necessity. This includes astronomy [117, 174, 217] and genomics [140].

These examples illustrate that nonsmooth functions play an integral role in modelling and exploiting structures of signals, for methods in signal processing.

Bilevel optimisation problems

To conclude this subsection, we introduce a class of optimisation problems in which there is another layer of complexity to account for. These are *bilevel optimisation problems*, in which one seeks to optimise the parameters that go into a variational regularisation problem. As mentioned earlier, it is common to parametrise regularisation terms, e.g. with ϑ in (1.8), yet choosing these parameters optimally can be difficult.

A bilevel optimisation problem is given by

$$\min_{\vartheta\in\Omega} E(x_\vartheta,\vartheta) \quad \text{s.t.} \quad x_\vartheta \in \arg\min_{x\in\mathbb{R}^n} F(x,\vartheta), \tag{1.11}$$

where $\Omega \subset \mathbb{R}^m$ is the parameter space, $E : \mathbb{R}^n \times \Omega \to \mathbb{R}$ is an *upper-level objective function* which takes as input the parameter choice ϑ as well as the corresponding signal reconstruction x_ϑ, and $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$ is a *lower-level objective function*, e.g. as given in (1.8).

What makes bilevel problems particularly difficult to solve is that the mapping $\vartheta \mapsto \arg\min_x F(x,\vartheta)$might be set-valued if $F(\cdot,\vartheta)$ is not strongly convex, nonsmooth if F is nonsmooth, and the mapping $\vartheta \mapsto E(x_\vartheta,\vartheta)$ is generally nonconvex due to the nonlinearity of the solution mapping $\vartheta \mapsto x_\vartheta$. Furthermore, computing x_ϑ for a single choice of ϑ can be computationally costly, and computing derivatives $D_\vartheta x_\vartheta$, when this is possible, even costlier.

In summary, bilevel problems for nonsmooth variational regularisation problems are nonsmooth, nonconvex, and computationally expensive. In fact, bilevel problems are in general NP hard [70]. See [40] for a review of bilevel optimisation for image processing.

The combination of nonsmoothness and nonconvexity has proven to be particularly challenging in optimisation problems, in comparison to for example convex, nonsmooth problems, whose theory is fairly well understood (see Section 2.4). With the exception of cases where the objective function can be split into a convex term and a smooth term, there are relatively few methods available for solving nonsmooth, nonconvex optimisation problems.

Several of the methods studied in this thesis address this class of problems, and are motivated by solving bilevel problems. That is, in Chapters 4 & 5, we propose derivative-free methods for nonsmooth, nonconvex optimisation problems, with applications to bilevel optimisation, and in Chapter 7, we study generalised differentiability of the solution mapping $\vartheta \mapsto x_\vartheta$ for nonsmooth variational problems.

We conclude these sections by reemphasising that nonsmoothness is an essential aspect of signal processing and optimisation methods. Yet in different contexts of optimisation, they continue to pose challenges, and hence there is a need for new optimisation methods that address these challenges, as well as new approaches to formulating and studying optimisation schemes. Furthermore, while explicit gradient descent and proximal mappings are powerful tools for solving various optimisation problems, they may not be applicable in cases where gradients are not accessible or the nonsmooth functions are not simple. It is therefore of interest to look at optimisation schemes that are applicable in more general settings.

1.1.3 Numerical integration

Numerical integration is the study of numerical methods for solving systems of differential equations. In recent years, this field has received increasing attention from the mathematical optimisation community, due to the idea that optimisation schemes can be understood through their relation to continuous-time differential systems and the numerical integration methods that connect them. We illustrate this with some examples.

The relevance of numerical integration to optimisation should not be surprising, considering that explicit and implicit gradient descent (1.2), (1.3), the two, main building blocks of first-order optimisation methods can be viewed as the forward and backward Euler method [114] respectively applied to the differential system known as the *gradient flow*,

$$\dot{x}(t) = -\nabla F(x(t)), \quad x(0) = x^0 \in \mathbb{R}^n, \quad t \geq 0. \tag{1.12}$$

Furthermore, their stability properties, e.g. unconditional dissipativity of (1.3) with respect to the time step τ_k, can be inferred from the properties of the Euler method.

A prominent example of the study of differential equations and numerical integration to address challenges in optimisation is that of understanding the *acceleration phenomenon*. In 1983, Nesterov introduced *accelerated gradient descent* [157] as a method that matches the optimal convergence rate of $O(1/k^2)$ for first-order methods on L-smooth, convex functions. Since the resurgence of first-order methods in the era of big data and high-dimensional optimisation, acceleration techniques have received significant attention in the past decade, for solving problems such as compressed sensing [16], training of deep and recurrent neural networks [210], and sparse linear regression [14].

In spite of its prevalence, the underlying dynamics of acceleration schemes are not well-understood, prompting several recent approaches to identify a framework in which to understand these schemes, taking perspectives from numerical integration. Su et al. [208] and Wibisono et al. [219] identify second-order ordinary differential equations (ODEs) that can be seen as continuous-time limits of the acceleration schemes. In the former case, this enables them to explain the oscillatory behaviour of acceleration scheme by interpreting the ODEs as damping systems. In the latter case, they present a family of *Bregman Lagrangian functionals* which generate the original and new acceleration schemes. Furthermore, they demonstrate that the choice of ODE discretisation method is crucial for whether the acceleration phenomena is retained in the iterative scheme.

Several works have contributed to this setting of numerical analysis of acceleration methods which bridges continous-time and discrete-time dynamics. Wilson et al. [220] approach this from the perspective of Lyapunov theory, presenting Lyapunov functions accounting for both continuous- and discrete-time dynamics. Betancourt et al. [23] present a framework of *sympletic optimisation*, i.e. considering perspectives of Hamiltonian dynamics and symplectic structure-preserving methods.

In a similar vein, recent papers by Maddison et al. [144] and França et al. [92] have studied conformal Hamiltonian systems, with the former focusing on how information about the the objective function's convex conjugate can be incorporated to obtain stronger convergence rates, and the latter on structure-preserving numerical methods and their relation to different iterative schemes.

Another central issue for iterative optimisation schemes is the choice of time step τ_k, which is closely tied to stability analysis of numerical methods. In this context, tools from numerical integration can be used to formulate iterative schemes that allow for the use of larger time steps and therefore faster progression towards the minimum. Eftekhari et al. [80] achieve this for strongly convex problems, by formulating explicit stabilised descent methods that use explicit Runge–Kutta methods to maximise the total length of time steps [1]. The theoretical analysis demonstrates robustness with respect to the objective function's condition number, and in numerical examples the method is shown to outperform accelerated gradient descent.

Other numerical integration methods include implicit Runge-Kutta methods, where energy dissipation is ensured under mild time step restrictions [104]. Finally we mention that one may consider gradient flows under non-Euclidean metrics, such as Bregman distances (see Chapter 6) and the Wasserstein metric [5, 200] (see Section 8.3.1).

Geometric numerical integration

In this thesis, we are particularly interested in geometric structure-preserving methods, which is the domain of *geometric numerical integration.* As described by Iserles & Quispel in *'Why Geometric Numerical Integration?'* [115], differential equations may exhibit geometric invariants, such as conservation laws of Hamiltonian energies, or Lie point symmetries, each of which imply that the solution to the differential equations is restricted to some lower-dimensional manifold. One is then interested in numerical methods which preserve these structures in some sense.

We highlight one class of methods from geometric integration, namely *discrete gradient methods* [97, 116, 148, 183]. These are designed for differential equations that can be written in *linear-gradient-form*, i.e.

$$\dot{x}(t) = A(x(t))\nabla F(x(t)), \tag{1.13}$$

where A is a matrix-valued function. By applying the chain rule, we derive

$$\frac{\mathrm{d}F(x(t))}{\mathrm{d}t} = \langle \nabla F(x(t)), A(x(t))\nabla F(x(t))\rangle,$$

from which one can observe that the system is conservative, i.e. F is constant along $x(t)$, if A is *skew-symmetric*, i.e. $A^* = -A$. Similarly, we observe that the system is dissipative if A is negative-definite i.e. $-A$ is positive-definite—see Section 2.3. In fact, [148, Proposition 2.1 & Proposition 2.8] show that conservative and dissipative systems can in general be expressed in linear-gradient form.

Discrete gradient methods preserve the geometric structures of linear-gradient systems, e.g. energy conservation and dissipation laws, as well as Lyapunov functions. Furthermore, the methods are *unconditionally stable*, in the sense that these properties are preserved for all discretisation time steps $\tau_k > 0$. This has prompted the study of discrete gradient methods applied to gradient flows for solving optimisation problems. We give some examples.

Grimm et al. [100] propose using discrete gradient methods for solving variational regularisation problems in image analysis. The applications include image inpainting and denoising, and they prove that for continuously differentiable objective functions, the methods converge to a set of stationary points. Furthermore, they compare the stability properties with other methods, such as Euler methods. In a similar setting, Ringholm et al. [185] consider the *Itoh–Abe* discrete gradient method for solving image inpainting problems regularised with Euler's elastica. These are nonconvex optimisation problems whose gradients are expensive to compute, while the Itoh–Abe discrete gradient (defined in Section 2.7) is derivative-free.

Their numerical results suggest that this method is competitive with state-of-art methods for nonconvex variational optimisation problems.

Going beyond gradient flows in Euclidean space, Celledoni et al. [44] extend the discrete gradient method to solve Riemannian gradient flow systems on manifolds. In this setting, they prove that the iterates of the method converge to a set of stationary points. They apply the method to solve eigenvalue problems, as well as imaging problems that can naturally be formulated on manifolds.

For some reviews of the field of geometric numerical integration, we refer the reader to [105, 115, 147].

In summary of this section, numerical integration has in recent years shown great promise in providing new perspectives and frameworks for addressing challenges in optimisation. As we detail in the next section, we are interested in various optimisation problems, including those concerning nonsmooth energies, and we are interested in the use of discrete gradient methods from geometric numerical integration applied in this setting.

1.2 Contributions

In what follows, we summarise the motivations for and contributions of each chapter.

Chapter 3: The foundations of discrete gradient methods for smooth optimisation

This chapter is based on the preprint [81], which is joint work done in collaboration with Matthias J. Ehrhardt, Torbjørn Ringholm, and Carola-Bibiane Schönlieb. The purpose of this chapter is to provide a comprehensive analysis of discrete gradient methods for the optimisation of continuously differentiable functions. While these optimisation methods have already been applied in various contexts for variational regularisation problems [100, 185], linear systems [153], and preserving Lyapunov functions [108], various aspects of the theoretical analysis have until now been lacking.

In this chapter, we address several issues, including convergence rates of the methods, well-posedness of the discrete gradient equation (2.8), and how to solve (2.8) efficiently. In particular, in Theorem 3.4 we prove for the three main discrete gradient methods that the discrete gradient equation admits a solution for all time steps. Furthermore, we prove that the discrete gradient methods essentially inherit the convergence rates of explicit gradient descent, yielding $O(1/k)$ rates for convex functions, and linear rates for strongly convex functions. Meanwhile, we propose a novel scheme for solving the discrete gradient equation, which we demonstrate to be theoretically and numerically superior in certain cases. Furthermore, we propose and study a natural generalisation of the Itoh–Abe discrete gradient method, akin to randomised coordinate descent and random pursuit methods. The theory is supported with numerical experiments.

Chapter 4: Discrete gradient methods for nonsmooth, nonconvex optimisation

This chapter is based on the preprint [184], which is joint work done in collaboration with Matthias J. Ehrhardt, G. R. W. Quispel, and Carola-Bibiane Schönlieb. In this chapter, we consider the Itoh–Abe discrete gradient method for solving nonsmooth, nonconvex optimisation problems. Since this discrete gradient is derivative-free, it provides us with a notion of gradient flow-type dissipation in a black-box setting where we only have access to function evaluations.

We consider the *Clarke subdifferential* framework [54], defined in Section 2.5, for locally Lipschitz continuous functions. In this setting, we prove for randomised extensions of the Itoh–Abe discrete gradient method, as well as deterministic variants, that the iterates converge to a limit set of Clarke stationary points. Convergence guarantees in the deterministic case is based on a property termed *cyclical density*. While the analysis in this chapter can be used for discrete gradient methods, they are immediately generalisable to other line search-based, derivative-free methods in the Clarke subdifferential setting, thus allowing for optimality analysis for a wider class of derivative-free optimisation algorithms. Noting that many bilevel problems are nonsmooth, nonconvex, and challenging to compute gradients for, we consider the proposed methods for solving these problems. Furthermore, we compare with state-of-art derivative-free optimisation algorithms, thereby demonstrating the competitiveness of the proposed methods.

Chapter 5: Discrete gradient methods for nonsmooth, nonconvex, constrained optimisation

In this chapter, we build on the analysis of the previous chapter for derivative-free optimisation of nonsmooth, nonconvex functions, by extending the algorithm and convergence analysis to constrained optimisation problems. The reason for this is that parameter-optimisation problems often involve constraints on the parameters, varying from explicit constraints to more complicated, implicitly defined constraints. We study this problem in a general setting, only assuming that the constraint is epi-Lipschitzian [188], which essentially means it is the level set of a locally Lipschitz continuous function. The Clarke subdifferential framework is extended to define stationary points constrained to a set, and in this framework, we prove that the algorithm converges to a set of stationary points.

Chapter 6: Bregman discrete gradient methods for sparse optimisation

This chapter is based on the article [20] published in the Journal of Mathematical Imaging and Vision, and which is joint work done in collaboration with Martin Benning and Carola-Bibiane Schönlieb. While in the previous chapters, we look at discrete gradient methods applied to gradient flows, in this chapter we consider discrete gradient methods applied to the inverse scale space flow [201], which is a dissipative differential system closely related to Bregman iterative methods. This system allows us to incorporate additional structure into the scheme, to promote sparsity or other features of the objective function and the ground truth. We study the Itoh–Abe discrete gradient method applied to this flow, and prove convergence in a nonsmooth, nonconvex subdifferential framework. We implement this method for different Bregman distances and objective functions, generalising well-known methods such as Gauss-Seidel and successive-over-relaxation (SOR) for sparse optimisation. Through numerical experiments, we observe that for sparse ground truths, the Bregman discrete gradient methods converge significantly faster than regular SOR. Furthermore, the analysis in this chapter opens the door for the application of discrete gradient methods to other, non-Euclidean gradient flows.

Chapter 7: Differentiation for nonsmooth bilevel optimisation

In this chapter, we focus exclusively on bilevel optimisation problems, seeking to exploit structured nonsmoothness of the corresponding variational problem to differentiate with respect to the parameters. To do so, we employ the framework of partial smoothness [130]. For a large class of bilevel problems, we demonstrate piecewise differentiability of the solution mapping in Theorem 7.29, allowing us to characterise the Clarke subdifferential of the bilevel objective function. Furthermore, we prove for various forward-backward type algorithms, including accelerated variants, that the algorithmic derivatives converge to the limiting, implicit derivative in Theorem 7.33.

1.3 Outline of chapters

Chapter 2: Mathematical preliminaries

In Chapter 2, we define notation and basic mathematical tools used throughout the thesis. Specifically, we provide background material for linear algebra, convex and nonconvex optimality analysis, tools for first-order optimisation methods, and finally discrete gradient methods and geometric numerical integration.

Chapter 3: The foundations of discrete gradient methods for smooth optimisation

After the introduction, we present a new existence result for the discrete gradient equation, based on the Brouwer fixed point theorem in Section 3.3. In Section 3.4 we study fixed point iterative methods for solving the discrete gradient equation, including a relaxed fixed point method with improved efficiency, while in Section 3.5, we study the dependence of the update $x^{k+1} \mapsfrom x^k$ on the time step $\tau_k > 0$ for the mean value discrete gradient and the Itoh–Abe methods. In Sections 3.6 and 3.7, we prove convergence rates for the discrete gradient methods, and convergence guarantees for functions that satisfy the strong Kurdyka–Łojasiewicz inequality, respectively. Before a brief discussion of preconditioned methods in Section 3.8, we present numerical results in Section 3.9.

Chapter 4: Discrete gradient methods for nonsmooth, nonconvex optimisation

In Section 4.1, we discuss the background for nonsmooth, nonconvex optimisation, and the purpose of the chapter. In Section 4.2 we provide the main theoretical results of the chapter, namely that the generalised Itoh–Abe methods converge to a connected set of Clarke stationary points. In Section 4.4 we propose an algorithm for solving black-box optimisation problems, and in Section 4.5 we provide numerical examples.

Chapter 5: Discrete gradient methods for nonsmooth, nonconvex, constrained optimisation

In Section 5.1 we propose a modification of the Itoh–Abe methods for constrained problems and discuss related works, while in Section 5.2 we provide preliminary results on epi-Lipschitzian sets and Clarke subdifferential analysis in the constrained setting. In Section 5.3

we study the proposed optimisation algorithm and prove that the methods converge to a connected set of Clarke stationary points in the constrained setting as well. In Section 5.4 we present numerical results.

Chapter 6: Bregman discrete gradient methods for sparse optimisation

After introducing the inverse scale space flow in Section 6.1, we propose to solve it using a Bregman discrete gradient method based on the ISS flow in Section 6.2. In Section 6.3 we prove well-posedness and convergence results for this method in a nonconvex, nonsmooth framework. Furthermore, in Sections 6.4 and 6.5, we discuss particular examples of Bregman discrete gradient methods and prove equivalencies between methods derived from different numerical integration schemes, respectively, before providing numerical results in Section 6.6.

Chapter 7: Differentiation for nonsmooth bilevel optimisation

In Section 7.1 we discuss bilevel optimisation of nonsmooth variational problems and motivations for studying the differential properties of the solution mapping. In Section 7.2, we review examples of nonsmooth bilevel problems and existing approaches for solving them in literature. In Section 7.3 we provide preliminary concepts for the subdifferential analysis, and prove for a sufficiently general class of variational problems that they are subdifferentially regular. In Section 7.4 we define partly smooth functions, and show under reasonable assumptions that the solution mapping is piecewise differentiable. In Section 7.5 we study algorithmic differentiation of various first-order methods for solving nonsmooth variational methods, and prove convergence guarantees to the implicit derivative. In Section 7.6 we present some numerical results, and in Section 7.7 we conclude.

Chapter 8: Conclusion & outlook

In Sections 8.1 and 8.2 we summarise and discuss the results of this thesis. In Section 8.3 we discuss future directions of research building on the work in this thesis. In particular, in Section 8.3.1, we consider solving the Wasserstein gradient flow with discrete gradients. In Section 8.3.2, we propose the use of mean value discrete gradient methods for nonsmooth objective functions, under assumptions of partial smoothness. In Section 8.3.3, we discuss future work for gradient-based approaches to bilevel problems, considering algorithmic differentiation of primal-dual methods, and studying stability of algorithmic differentiation methods when the number of iterations is determined by a stopping rule.

Chapter 2

Mathematical preliminaries

In this section, we provide mathematical preliminaries which will be used throughout the thesis. We first consider basic properties of differentiable functions, followed by theory of the class of convex, proper, lower semicontinuous functions. Next we provide an overview of nonconvex generalised differential theory, which in comparison to its convex counterpart is rather less unified. We conclude with an overview of geometric numerical integration and in particular discrete gradients.

2.1 Basic notation and conventions

In a Euclidean space setting, we denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the norm and associated inner product. For $x \in \mathbb{R}^n$ and $p > 0$, the ℓ^p-norm $\|\cdot\|_p$ is defined as

$$\|x\|_p := \sqrt[p]{x_1^p + x_2^p + \ldots + x_n^p},$$

while $\|x\|_\infty := \max_{i=1,\ldots,n} |x_i|$, and $\|x\|_0 := |\operatorname{supp}(x)|$.

We denote by $(e^i)_{i=1}^n$ the standard coordinate vectors in $\mathbb{R}^n$. We denote by $[x,y]$ the line segment between two points $x, y \in \mathbb{R}^n$ i.e.

$$[x,y] := \left\{\lambda x + (1-\lambda)y \,:\, \lambda \in [0,1]\right\}.$$

For $\varepsilon > 0$, $x \in \mathbb{R}^n$, we denote by $B_\varepsilon(x)$ the *open ball of radius* ε *at* x, $\{y \in \mathbb{R}^n \,:\, \|x-y\| < \varepsilon\}$, and by $\overline{B}_\varepsilon(x)$ the *closed ball of radius* ε *at* x, $\{y \in \mathbb{R}^n \,:\, \|x-y\| \leq \varepsilon\}$. We denote by S^{n-1} the *unit sphere in* $\mathbb{R}^n$, $\{x \in \mathbb{R}^n \,:\, \|x\| = 1\}$.

We summarise *big* and *small o-notation*. For two functions $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^l$, if $\|f(x)\|/\|g(x)\| \to 0$ as $\|x\| \to 0$, then $f(x) = o(g(x))$. If there is $\varepsilon > 0$ and $C > 0$ such that $\|x\| < \varepsilon$ implies $\|f(x)\| \le C\|g(x)\|$, then $f(x) = O(g(x))$.

Similarly, suppose $(x^n)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ and $(y^n)_{k\in\mathbb{N}} \subset \mathbb{R}^m$ are two sequences. If $\|x^k\|/\|y^k\| \to 0$ as $k \to \infty$, then $x^k = o(y^k)$, while if there is $C > 0$ and $K \in \mathbb{N}$ such that $\|x^k\| \le C\|y^k\|$ for all $k \ge K$, then $x^k = O(y^k)$.

2.2 Notation and results for functions

First we introduce some notation for differentiable functions. For $k \in \mathbb{N}$, the set of k-times continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}^m$ is denoted by $C^k(\mathbb{R}^n;\mathbb{R}^n)$, and $C^k(\mathbb{R}^n)$ for short when $m = 1$. For $f \in C^1(\mathbb{R}^n;\mathbb{R}^m)$, we denote by $\nabla f(x)$ its gradient at x, and for $f \in C^2(\mathbb{R}^n)$, we denote by $\nabla^2 f(x)$ its Hessian at x. Furthermore, for parametrised functions $f(x,\vartheta)$ we use D instead of ∇ to denote differentiation with respect to the parameters ϑ, and we write $\nabla_x f(x,\vartheta)$ and $D_\vartheta f(x,\vartheta)$ to denote differentiation with respect to first and second argument respectively.

We say that a function f on $\mathbb{R}^n$ is *set-valued* in $\mathbb{R}^m$ if for each $x \in \mathbb{R}^n$, $f(x)$ is a subset of $\mathbb{R}^m$, and we write $f : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$. An important example of a set-valued function is the subdifferential given in Section 2.4.

Definition 2.1 (Graph). *For a function $f : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, its* graph *is the subset of $\mathbb{R}^n \times \mathbb{R}^m$ given by*

$$\operatorname{gph} f := \{(x,v) \in \mathbb{R}^n \times \mathbb{R}^m \ : \ v \in f(x)\}.$$

For single-valued functions, the graph is defined analogously.

Definition 2.2 (Support). *For a function $f : \mathbb{R}^n \to \mathbb{R}^m$, its support is the set of points for which the function does not vanish, i.e.*

$$\operatorname{supp} f := \{x \in \mathbb{R}^n \ : \ f(x) \neq 0\}.$$

We also state the implicit function theorem, which we will apply to study properties of discrete gradient equations in Chapter 3, as well as for computing gradients for bilevel problems in Chapter 7.

Proposition 2.3 (Implicit function theorem [22, Proposition A.25]). *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a function such that for some $(x^*,y^*) \in \mathbb{R}^n \times \mathbb{R}^m$, $f(x^*,y^*) = 0$, f is locally C^1-smooth around (x^*,y^*), and $\nabla_x f(x^*,y^*) \in \mathbb{R}^{n\times n}$ is a nonsingular matrix. Then there are neighbourhoods X*

and Y of x^* and y^* and a continuous function $\phi : Y \to X$, such that

$$f(x,y) = 0 \quad \Longleftrightarrow \quad x = \phi(y), \qquad \textit{for all } x \in X, y \in Y.$$

Furthermore, if f is C^p-smooth for $p \in \mathbb{N}$, then ϕ is C^p-smooth with gradient

$$\nabla\phi(y) = -(\nabla_x f(\phi(y),y))^{-1}\nabla_y f(\phi(y),y).$$

2.3 Linear algebra

For a matrix $A \in \mathbb{C}^{n,n}$, denote by $a^i \in \mathbb{R}^n$ its ith row, A^H its Hermitian and A^* its adjoint. A matrix is said to be *Hermitian* (resp. *self-adjoint*) if $A^H = A$ (resp. $A^* = A$).

We denote by $\ker A$ the *kernel* of A, i.e. the subspace of vectors x such that $Ax = 0$.

Consider a Hermitian matrix $A \in \mathbb{C}^{n,n}$. We say that A is *positive-definite* if

$$\langle x, Ax \rangle > 0 \quad \text{for all } x \in \mathbb{C}^n,$$

and *positive-semidefinite* if

$$\langle x, Ax \rangle \geq 0 \quad \text{for all } x \in \mathbb{C}^n.$$

A positive-definite matrix is always nonsingular, i.e. it admits an inverse $A^{-1} \in \mathbb{C}^{n,n}$, which is also positive-definite.

We denote by I_n the identity matrix in $\mathbb{R}^{n,n}$ and by 0_n the zero matrix. When the dimension is unambiguous, we occasionally write I instead.

The (operator) norm of a matrix $A \in \mathbb{R}^{m,n}$ is defined as

$$\|A\| := \sup_{x \in S^{n-1}} \|Ax\|,$$

while the Frobenius norm is defined as

$$\|A\|_F := \sqrt{\sum_{i,j=1}^{n} a_{i,j}^2},$$

where $a_{i,j} = a^i_j$.

The *rank* of a matrix A is the dimension of the column space (or equivalently row space) of the matrix, i.e. the number of linearly independent column vectors, and is denoted by $\mathrm{rank}(A)$.

Definition 2.4 (Spectrum). *The* spectrum *of a square matrix $A \in \mathbb{C}^{n,n}$ is its set of eigenvalues,*

$$\sigma(A) := \{\lambda \in \mathbb{C} \,:\, \exists x \in \mathbb{C}^n \setminus \{0\} \text{ with } \lambda x = Ax\}.$$

For a positive-definite matrix A, its *condition number* is the ratio $\kappa_A := \lambda_n/\lambda_1 \geq 1$, where λ_n and λ_1 are respectively the matrix' largest and smallest eigenvalues. A large condition number means the matrix is *ill-conditioned* while a low condition number means it is *well-conditioned.*

Definition 2.5 (Spectral radius). *The* spectral radius *of a square matrix $A \in \mathbb{C}^{n,n}$ is*

$$\rho(A) := \sup_{\lambda \in \sigma(A)} |\lambda|.$$

Gelfand's formula denotes an important relationship between A and its spectral radius.

Proposition 2.6 (Gelfand's formula [124, Theorem 7.5.5]). *For any square matrix $A \in \mathbb{C}^{n,n}$, the following limit holds:*

$$\lim_{k \to \infty} \sqrt[k]{\|A^k\|} = \rho(A).$$

The following result is a variation on [179, Chapter 2, Theorem 1], a key ingredient for showing convergence of various iterative schemes, based on Gelfand's formula.

Proposition 2.7. *Let $(A_k)_{k\in\mathbb{N}} \subset \mathbb{C}^{n,n}$, $(b^k)_{k\in\mathbb{N}} \in \mathbb{C}^n$, and for $d^0 \in \mathbb{C}^n$, define the iterates*

$$d^{k+1} = A_k d^k + b^k, \quad k \in \mathbb{N}.$$

If $A_k \to A$ and $b^k \to b$ with $\rho(A) < 1$, then the iterates $(d^k)_{k\in\mathbb{N}}$ converge linearly to the fixed point $(I - A)^{-1}b$.

Proof. Since $\rho(A) < 1$, it follows from Gelfland's formula that the Neumann sequence $\sum_{i=0}^{\infty} A^i$ is well-defined and equal to $(I-A)^{-1}$. Therefore, $d^* := (I-A)^{-1}b$ is the unique fixed point of the mapping $d \mapsto Ad + b$.

Write $y^k = d^k - d^*$, $r^k = (A_k - A)y^k$, and $s^k = (A_k - A)d^* + b^k - b$. Then

$$y^{k+1} = Ay^k + r^k + s^k, \qquad r^k = o(y^k), \qquad \|s^k\| \to 0.$$

Thus

$$
\begin{aligned}
y^{k+1} &= A^{k+1}y^0 + \sum_{i=1}^{k} A^{k-i}r^i + \sum_{i=1}^{k} A^{k-i}s^i,\\
\|y^{k+1}\| &\le \|A^{k+1}\|\|y^0\| + \sum_{i=1}^{k} \|A^{k-i}\|\|r^i\| + \sum_{i=1}^{k} \|A^{k-i}\|\|s^i\|.
\end{aligned}
$$

By [179, Chapter 2, Lemma 1], for any $\rho \in (\rho(A), 1)$ there is $c > 0$ such that $\|A^k\| \le c\rho^k$ for all $k \in \mathbb{N}$, which implies

$$\|y^{k+1}\| \le c\rho^{k+1}\|y^0\| + c\sum_{i=1}^{k} \rho^{k-i}\|r^i\| + c\sum_{i=1}^{k} \rho^{k-i}\|s^i\|.$$

Since $\|s^i\| \to 0$, the third term vanishes as $k \to \infty$. Finally, as $r^k = o(y^k)$, the result follows. □

In Chapter 7, we will repeatedly make use of the following result to simplify the analysis.

Proposition 2.8. *If A and $B \in \mathbb{C}^{n,n}$ are self-adjoint matrices, and B is positive-definite, then $\sigma(AB) = \sigma(BA) \subset \mathbb{R}$.*

Proof. By [124, Theorem 9.4.2], there is a self-adjoint, positive-definite square root of B, $\sqrt{B} \in \mathbb{C}^{n,n}$, such that $\sqrt{B}^2 = B$. One can verify that $\sigma(AB) = \sigma(\sqrt{B}A\sqrt{B})$. It remains to note that $\sqrt{B}A\sqrt{B}$ is self-adjoint, so by [124, Theorem 9.2.1], $\sigma(\sqrt{B}A\sqrt{B}) \subset \mathbb{R}$.

To show the equality $\sigma(AB) = \sigma(BA)$, note that for any square matrix A, the eigenvalues of A^H equal the complex conjugates of the eigenvalues of A. Then the equality follows from the fact that $(AB)^H = BA$ and that $\sigma(AB) \subset \mathbb{R}$. □

Finally, we introduce the *(Moore–Penrose) pseudoinverse* which can be defined accordingly in finite-dimensional spaces.

Definition 2.9 (Moore–Penrose pseudoinverse). *Let $A \in \mathbb{R}^{m,n}$ be a matrix. The* Moore–Penrose pseudoinverse *of A, $A^\dagger \in \mathbb{R}^{n,m}$, is the (unique) matrix which satisfies the following four conditions.*

$$AA^\dagger A = A, \quad A^\dagger AA^\dagger = A^\dagger, \quad (AA^\dagger)^* = AA^\dagger, \quad (A^\dagger A)^* = A^\dagger A.$$

For square, invertible matrices A, we have $A^\dagger = A^{-1}$. However, pseudoinverses are also uniquely defined for non-square and singular matrices. For further details on the pseudoinverse, see [83, Section 2.1].

2.4 Convex analysis

We now review basic results of convex analysis. For the theory of convex functions and their subdifferentials, see [82, 112, 191].

Definition 2.10 (Convex sets and functions). *A set $C \subset \mathbb{R}^n$ is* convex *if for all $x, y \in C$ and $\lambda \in (0,1)$, one has $\lambda x + (1-\lambda)y \in C$.*

A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex *if for all $x, y \in \mathbb{R}^n$ and $\lambda \in (0,1)$, one has*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

Here $\overline{\mathbb{R}}$ is the *extended real number line* $\mathbb{R} \cup \{\pm\infty\}$.

Definition 2.11 (Lower semicontinuity). *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* lower semicontinuous at *$x \in \mathbb{R}^n$ if for all sequences $(x^k)_{k\in\mathbb{N}}$ converging to x, one has*

$$\liminf_{k\to\infty} f(x^k) \leq f(x).$$

If this holds at all $x \in \mathbb{R}$, then we say that f is lower semicontinuous.

Definition 2.12 (Effective domain). *The* effective domain *of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined as* $\mathrm{dom}(f) = \{x \in \mathbb{R}^n \ : \ f(x) < \infty\}$.

We call a function *proper* if $\mathrm{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. The following set of functions is central to convex analysis.

Definition 2.13. *The set $\Gamma_0(\mathbb{R}^n)$ consists of all functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ that are convex, proper, and lower semicontinuous.*

Definition 2.14 (Subgradients and subdifferentials). *For a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $p \in \mathbb{R}^n$ is a* subgradient of f at x *if*

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

The subdifferential of f at x *is the set $\partial f(x)$ of all subgradients of f at x.*

Remark 2.15. *We use the same subdifferential notation for the* general *subdifferential in Chapter 7, which generalises the subdifferential to nonconvex functions.*

The following property is immediate and generalises first-order optimality conditions to nonsmooth, convex functions.

Proposition 2.16. *If $f \in \Gamma_0(\mathbb{R}^n)$, then $x \in \mathbb{R}^n$ is a minimiser of f if and only if $0 \in \partial f(x)$.*

Definition 2.17 (Strong convexity)**.** *A proper, convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* strongly convex with constant $\mu > 0$*, or* μ-convex *for short, if either of the following (equivalent) conditions hold.*

(i) The function $f(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ is convex.

(ii) $f\big(\lambda x + (1-\lambda)y\big) \le \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2$ for all x, y in $\mathbb{R}^n$ and $\lambda \in [0,1]$.

(iii) For all $x, y \in \operatorname{dom} f$, $p \in \partial f(x)$, and $q \in \partial f(y)$, one has $\langle p-q, x-y\rangle \ge \mu\|x-y\|^2$.

Remark 2.18. *In this context, 0-convexity simply means convexity.*

The following property can be derived from the third characterisation of μ-convexity above.

Proposition 2.19. *If $f : \mathbb{R}^n \to \mathbb{R}$ is C^2-smooth and μ-convex for $\mu > 0$, then for all $x \in \mathbb{R}^n$, the Hessian $\nabla^2 f(x)$ is positive-definite with*

$$\langle y, \nabla^2 f(x) y\rangle \ge \mu\|y\|^2.$$

An important class of functions in $\Gamma_0(\mathbb{R}^n)$ are *indicator functions* of convex sets, defined in (1.6). Since

$$\operatorname*{arg\,min}_{x\in C} f(x) = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} f(x) + \delta_C(x),$$

we can thus treat both constrained and unconstrained, convex optimisation problems under the same framework.

We have already mentioned the proximal mapping in Chapter 1.

Definition 2.20 (Proximal mappings)**.** *For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}} \in \Gamma_0(\mathbb{R}^n)$ and parameter $\lambda > 0$, the* proximal mapping *is the function*

$$\operatorname{prox}_{\lambda f} : \mathbb{R}^n \to \mathbb{R}^n, \quad \operatorname{prox}_{\lambda f}(x) := \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \lambda f(y) + \frac{1}{2}\|y-x\|^2.$$

We also introduce another important concept in convex analysis, namely *convex conjugates.*

Definition 2.21 (Convex conjugate)**.** *For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, its* convex conjugate *is the function $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ given by*

$$f^*(y) = \sup_{x\in\mathbb{R}^n} \langle y, x\rangle - f(x).$$

While we do not go into the details of this, convex conjugates are important for primal-dual methods and for transforming a variational optimisation problem into a saddle-point problem. We point out the following result.

Proposition 2.22 ([192, Theorem 11.1])**.** *If $f \in \Gamma_0(\mathbb{R}^n)$, then $f^* \in \Gamma_0(\mathbb{R}^n)$, and $(f^*)^* = f$.*

We also cover (generalised) Bregman distances [30], an important concept for convex analysis and optimisation methods.

Definition 2.23 (Bregman distance)**.** *For a function $f \in \Gamma_0(\mathbb{R}^n)$, $x \in \operatorname{dom} f$ and $p \in \partial f(x)$, the* Bregman distance of f *is the function*

$$D_f^p(y,x) := f(y) - f(x) - \langle p, y-x \rangle.$$

Remark 2.24. *It follows from the definition of subgradients that $D_f^p(y,x) \geq 0$ for all y, and that if f is μ-convex, then $D_f^p(y,x) \geq \frac{\mu}{2}\|y-x\|^2$.*

Example 2.25. *If $f(x) = \|x\|^2/2$, then $D_f^p(y,x) = \|y-x\|^2/2$, i.e. the Euclidean norm squared.*

While Bregman distances are nonnegative, they are not metrics as they generally do not satisfy symmetry or a triangle inequality. However, we can induce symmetry accordingly.

Definition 2.26 (Symmetric Bregman distance)**.** *Given $p \in \partial f(x)$ and $q \in \partial f(y)$, the* symmetric Bregman distance between x and y *is given by*

$$D_f^{\text{symm}}(x,y) = D_f^q(y,x) + D_f^p(x,y) = \langle q-p, y-x \rangle.$$

Although we drop the subgradient superscript in D_f^{symm}, the choice of subgradient will be clear from the context.

2.5 Nonconvex subdifferential theory

While convex optimality analysis is a well-understood and compact area, the picture is quite different for *nonconvex*, nonsmooth optimality analysis. In this setting, there are various generalisations of the subdifferential for different classes of functions, and there exist research surveys dedicated merely to mapping the differences and nuances between these generalisations. See e.g. [25, 26], and see [69] for various notions of stationarity in the context of bilevel optimisation.

In this thesis, we mainly focus on the nonconvex, nonsmooth optimality analysis framework [54] proposed by Francis H. Clarke in his doctoral thesis [53], now termed the Clarke subdifferential framework. It generalises the gradient of a differentiable function, as well as the subdifferential [82] of a convex function.

There are alternative frameworks for generalising differentiability of nonsmooth, nonconvex functions, each with different analytical properties. For example, the Michel–Penot subdifferential [152] coincides with the Gâteaux derivative when this exists, unlike the Clarke subdifferential, which is larger and only coincides with strict derivatives [95]. However, the Clarke subdifferential is outer semicontinuous, making it in most cases the preferred framework for analysis. See [25] by Borwein and Zhu for a survey of various subdifferentials, published on the 25th birthday of the Clarke subdifferential.

2.5.1 Clarke subdifferential theory

We summarise the main concepts of the Clarke subdifferential for locally Lipschitz continuous, nonsmooth, nonconvex functions $f : \mathbb{R}^n \to \mathbb{R}$, and refer to [54] for further details.

We first define *local Lipschitz continuity.*

Definition 2.27 (Lipschitz continuity). *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is* locally Lipschitz continuous near *x with Lipschitz constant $L > 0$ if there is a neighbourhood N_x of x such that for all $y, z \in N_x$, one has*

$$\|f(y) - f(z)\| \leq L\|y - z\|.$$

f is locally Lipschitz continuous *if the above property holds for all $x \in \mathbb{R}^n$.*

Definition 2.28 (Clarke directional derivative). *For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$, f is* Clarke directionally differentiable *at x along d if the limit*

$$f^o(x;d) := \limsup_{y \to x,\, \lambda \downarrow 0} \frac{f(y + \lambda d) - f(y)}{\lambda}$$

exists. If so, $f^o(x;d)$ is called the Clarke directional derivative.

Remark 2.29. *Locally Lipschitz continuous functions are Clarke directionally differentiable.*

Clarke directional differentiability extends directional differentiability, which we define here for completeness.

Definition 2.30 (Directional derivative). *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* directionally differentiable at $x \in \mathbb{R}^n$ along $d \in \mathbb{R}^n$ *if the limit*

$$f'(x;d) := \lim_{\lambda \downarrow 0} \frac{f(x+\lambda d) - f(x)}{\lambda}$$

exists. We refer to $f'(x;d)$ as the directional derivative of f at x along d.

Definition 2.31 (Clarke subdifferential). *For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $x \in \mathbb{R}^n$, $p \in \mathbb{R}^n$ is a* Clarke subgradient of f at x *if*

$$f^o(x;d) \geq \langle d, p \rangle \text{ for all } d \in \mathbb{R}^n.$$

The Clarke subdifferential of f at x, *denoted by $\partial^C f(x)$, is the set of all such subgradients.*

The Clarke subdifferential is well-defined for locally Lipschitz functions and coincides with the standard subdifferential for convex functions [54, Proposition 2.2.7]. Furthermore, if f is strictly differentiable at $x \in \mathbb{R}^n$, then $\partial^C f(x) = \{\nabla f(x)\}$ [54, Proposition 2.2.4]. We additionally state three useful results, all of which can be found in [54, Chapter 2].

Proposition 2.32. *Suppose $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is locally Lipschitz continuous near $x \in \mathbb{R}^n$ with Lipschitz constant L. Then*

(i) $\partial^C f(x)$ is nonempty, convex and compact, and $\partial^C f(x) \subseteq B_L(0)$.

(ii) $\partial^C f(x)$ is outer semicontinuous. That is, for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\partial^C f(y) \subset \partial^C f(x) + B_\varepsilon(0), \quad \text{for all } y \in B_\delta(x).$$

(iii) Denote by $\mathcal{D}(f)$ the set of points $x \in \mathbb{R}^n$ at which f is differentiable. Then

$$\partial^C f(x) = \mathrm{co}\left\{ d \in \mathbb{R}^n : \exists (x^k)_{k \in \mathbb{N}} \subset \mathcal{D}(f) \text{ s.t. } x^k \to x \text{ and } \nabla f(x^k) \to d \right\}. \tag{2.1}$$

Here co *refers to the convex hull of the set.*

Similarly to the convex case, the Clarke subdifferential framework provides us with the following notion of a first-order optimality condition for nonsmooth, nonconvex functions.

Definition 2.33 (Clarke stationary point). *For $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $x^* \in \mathbb{R}^n$ is a* Clarke stationary point of f *if* $0 \in \partial^C f(x^*)$.

For our purposes, we also define Clarke directional stationarity.

Definition 2.34 (Directional Clarke stationarity)**.** *For a direction $d \in \mathbb{R}^n \setminus \{0\}$, we say that $f : \mathbb{R}^n \to \mathbb{R}$ is* Clarke directionally stationary at x^* along d *if*

$$\min\left\{f^o(x^*;d), f^o(x^*;-d)\right\} \geq 0.$$

Remark 2.35. *A point x^* is Clarke stationary if and only if f is Clarke directionally stationary at x^* along d for all $d \in S^{n-1}$.*

Any local maxima and minima of a function are Clarke stationary points. If f is convex, then stationary points coincide with the global minima, by Proposition 2.16. For more general classes of functions, the concept of Clarke stationary points also reduces to convex, first-order optimality conditions.

Definition 2.36 (Pseudoconvexity [175])**.** *A locally Lipschitz continuous function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* pseudoconvex *if, for all $x, y \in \mathbb{R}^n$,*

$$f(y) < f(x) \implies \forall p \in \partial^C f(x), \quad \langle p, y - x\rangle < 0.$$

Remark 2.37. *If f is pseudoconvex, then any Clarke stationary point is a global minimum [11].*

2.6 First-order optimisation methods

We now return to first-order optimisation methods, which we categorise as methods that make use of gradients, subgradients, and proximal mappings.

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function. A crucial subclass of these methods are *L-smooth functions*.

Definition 2.38 (L-smooth)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* L-smooth *for $L > 0$ if it is continuously differentiable and the gradient is Lipschitz continuous with Lipschitz constant L.*

We state some properties of L-smooth functions.

Proposition 2.39. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. Then for all $x, y \in \mathbb{R}^n$, the following holds.*

(i) $f(y) - f(x) \leq \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$. (Descent lemma)

(ii) $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y) - \frac{\lambda(1-\lambda)L}{2}\|x - y\|^2$ *for all* $\lambda \in [0,1]$.

(iii) If f is furthermore C^2-smooth, then $\|\nabla^2 f(x)\| \leq L$ for all x.

Proof. Property *(i)*. [22, Proposition A.24].

Property *(ii)*. It follows from property *(i)* that the function $x \mapsto \frac{L}{2}\|x\|^2 - f(x)$ is convex, which in turn yields the desired inequality.

Property *(iii)*. This follows from [158, Lemma 1.2.2] □

For μ-convex, L-smooth functions $f : \mathbb{R}^n \to \mathbb{R}$, its *condition number* is $\kappa_f := L/\mu$, and we use the same terminology as for a matrix' condition number. Note furthermore by Proposition 2.39 and Proposition 2.19 that κ_f is also an upper bound for the condition number of the Hessian of f.

Recall the explicit gradient descent method (1.2) for $x^k \in \mathbb{R}^n$ and time step τ_k. Observe that the update for x^{k+1} can be expressed as the solution to the variational problem given by

$$x^{k+1} = \operatorname*{arg\,min}_{y\in\mathbb{R}^n} f(x^k) + \langle y - x^k, \nabla f(x^k)\rangle + \frac{1}{2\tau_k}\|y - x^k\|^2. \tag{2.2}$$

Then, by Proposition 2.39, if f is L-smooth and $\tau_k = \sigma/L$ for $\sigma \in (0,2)$, one has

$$f(x^{k+1}) \leq f(x^k) - \frac{2\sigma - \sigma^2}{2L}\|\nabla f(x^k)\|^2 \leq f(x^k),$$

i.e. for $\tau_k \in (0, 2/L)$, the scheme is dissipative.

Similarly, recall that updates for implicit gradient descent also can be defined via a minimisation step, i.e. (1.4), and that this update is unconditionally dissipative with respect to τ_k.

Furthermore, note that in each case, the time step has a new interpretation as the weighting for the Euclidean distance term $\|y - x^k\|^2$. With this in mind, one can alter the descent scheme by replacing the Euclidean energy with a different measure of distance. Of particular relevance for optimisation is the use of Bregman distances.

For a function $J \in \Gamma_0(\mathbb{R}^n)$, time step $\tau_k > 0$, current iterate $x^k \in \mathbb{R}^n$, and subgradient iterate $p^k \in \partial J(x^k)$, the *Bregman (proximal minimisation) method* [18] is given by

$$x^{k+1} = \operatorname*{arg\,min}_{y\in\mathbb{R}^n} f(y) + \frac{1}{\tau_k} D_J^{p^k}(y, x^k). \tag{2.3}$$

Similarly, the *linearised Bregman method* is given by

$$x^{k+1} = \operatorname*{arg\,min}_{y\in\mathbb{R}^n} f(x^k) + \langle y - x^k, \nabla f(x^k)\rangle + \frac{1}{\tau_k} D_J^{p^k}(y, x^k). \tag{2.4}$$

Note that one requires strong convexity, or at least strict convexity, of J for the linearised Bregman method to be well-defined. In [13], Beck & Teboulle illustrate that the popular *mirror descent* algorithm [156] can be rewritten as a linearised Bregman method.

We highlight this interpretation of first-order descent methods for two reasons. First, Bregman iterations appear in several contexts in this thesis, first as a motivation for introducing Bregman discrete gradient methods in Chapter 6, and in the study of algorithmic differentiation in Chapter 7. Second, this interpretation provides a way of defining gradient flows with respect to non-Euclidean energies, via so-called *minimising movements schemes* [5], an idea we return to in Chapter 6 and Section 8.3.1.

See [212] for a recent review of first-order optimisation methods, with a focus on Bregman iterations. We furthermore refer the reader to [104] for a review of various energy-diminishing discretisation methods for gradient systems, including implicit Euler and discrete gradient methods.

2.7 Geometric numerical integration and discrete gradients

In Section 1.1.3, we discussed numerical integration and geometric numerical integration, and their applications to optimisation. In what follows we define discrete gradients, introduce the three most common examples of discrete gradients, and consider their applicability to the Euclidean gradient flow (1.12).

Definition 2.40 (Discrete gradient). *Let f be a continuously differentiable function. A* discrete gradient *is a continuous map $\overline{\nabla} f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that for all $x, y \in \mathbb{R}^n$,*

$$\langle \overline{\nabla} f(x,y), y-x \rangle = f(y) - f(x) \quad \textit{(Mean value)}, \tag{2.5}$$

$$\lim_{y \to x} \overline{\nabla} f(x,y) = \nabla f(x) \qquad \textit{(Consistency)}. \tag{2.6}$$

Before we present the discrete gradient method, we briefly consider the dissipative structure of the gradient flow (1.12). By applying the chain rule, we compute

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\|\nabla f(x(t))\|^2 = -\|\dot{x}(t)\|^2 \leq 0. \tag{2.7}$$

Thus the gradient flow is characterised by the decrease of $f(x(t))$ along $x(t)$ at the rate of $\|\nabla f\|^2$ or equivalently $\|\dot{x}\|^2$.

We now introduce the discrete gradient method for optimisation. For $x^0 \in \mathbb{R}^n$ and time steps $(\tau_k)_{k\in\mathbb{N}} \subset (0,+\infty)$, we solve

$$x^{k+1} = x^k - \tau_k \overline{\nabla} f(x^k, x^{k+1}). \tag{2.8}$$

This scheme preserves the dissipative structure of gradient flows, as can be seen by applying (2.5),

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= \langle \overline{\nabla} f(x^k, x^{k+1}), x^{k+1} - x^k \rangle \\ &= -\tau_k \|\overline{\nabla} f(x^k, x^{k+1})\|^2 = -\tau_k \| \frac{x^{k+1} - x^k}{\tau_k} \|^2. \end{aligned} \tag{2.9}$$

Note that the decrease holds for all time steps $\tau_k > 0$, and that (2.9) can be seen as a discrete analogue of the dissipative structure of gradient flows (2.7), replacing derivatives by finite differences.

We assume throughout the thesis that there are bounds $\tau_{\max} \geq \tau_{\min} > 0$ such that for all $k \in \mathbb{N}$,

$$\tau_{\min} \leq \tau_k \leq \tau_{\max}. \tag{2.10}$$

While there are infinitely many discrete gradients, there are three constructions that are of particular relevance. We state these here.

1. *The Gonzalez discrete gradient* [97] (also known as the midpoint discrete gradient) is given by

$$\overline{\nabla} f(x,y) = \nabla f\left(\frac{x+y}{2}\right) + \frac{f(y) - f(x) - \langle \nabla f(\frac{x+y}{2}), y-x \rangle}{\|x-y\|^2}(y-x), \quad x \neq y. \tag{2.11}$$

   This discrete gradient was introduced by Oscar Gonzalez in 1996, with the aim of providing a formalistic way of numerically solving Hamiltonian systems.

2. *The mean value discrete gradient* [106], used for example in the average vector field method [45], is given by

$$\overline{\nabla} f(x,y) = \int_0^1 \nabla f\big((1-s)x + sy\big)\,\mathrm{d}s, \tag{2.12}$$

   where $\int$ denotes integration.

3. *The Itoh–Abe discrete gradient* [116] (also known as the coordinate increment discrete gradient) is given by

$$\overline{\nabla} f(x,y) = \begin{pmatrix} \frac{f(y_1,x_2,\ldots,x_n)-f(x)}{y_1-x_1} \\ \frac{f(y_1,y_2,x_3,\ldots,x_n)-f(y_1,x_2,\ldots,x_n)}{y_2-x_2} \\ \vdots \\ \frac{f(y)-f(y_1,\ldots,y_{n-1},x_n)}{y_n-x_n} \end{pmatrix}, \tag{2.13}$$

where $0/0$ is interpreted as $[\nabla f(x)]_i$.

Proposition 2.41. *The mappings defined by* (2.11)-(2.13) *are discrete gradients.*

Proof. Continuity of the mappings follows from continuous differentiability of the function f.

The mean value property (2.5) is straightforward to verify for the Gonzalez and Itoh–Abe discrete gradients, by plugging in their respective expressions. For the mean value discrete gradient, we derive

$$\left\langle \int_0^1 \nabla f\left((1-s)x+sy\right) \mathrm{d}s, y-x \right\rangle = \int_0^1 \langle \nabla f\left((1-s)x+sy\right), y-x\rangle \,\mathrm{d}s = f(y)-f(x),$$

where the final equality follows by applying the fundamental theorem of calculus [198, Theorem 7.16] to the function $g(s) := f((1-s)x+sy)$.

Finally, as with continuity of the mappings, the consistency property (2.6) can be verified directly using continuous differentiability of f. □

While the first two discrete gradients are gradient-based , the Itoh–Abe discrete gradient is derivative-free, and is evaluated by computing successive, coordinate-wise difference quotients. In an optimisation setting, the Itoh–Abe discrete gradient is often preferable to the others, as it is relatively computationally inexpensive. Solving the implicit equation (2.8) with this discrete gradient amounts to successively solving n scalar equations of the form

$$\begin{aligned}
x_1^{k+1} &= x_1^k - \tau_k \frac{f(x_1^{k+1}, x_2^k, \ldots, x_n^k) - f(x^k)}{x_1^{k+1} - x_1^k} \\
x_2^{k+1} &= x_2^k - \tau_k \frac{f(x_1^{k+1}, x_2^{k+1}, x_3^k, \ldots, x_n^k) - f(x_1^{k+1}, x_2^k, \ldots, x_n^k)}{x_2^{k+1} - x_2^k} \\
&\vdots \\
x_n^{k+1} &= x_n^k - \tau_k \frac{f(x^{k+1}) - f(x_1^{k+1}, x_2^{k+1}, \ldots, x_n^{k+1}, x_n^k)}{x_n^{k+1} - x_n^k}.
\end{aligned}$$

Chapter 3

The foundations of discrete gradient methods for smooth optimisation

3.1 Introduction

This chapter is based on the preprint [81] and is joint work with Matthias J. Ehrhardt, Torbjørn Ringholm, and Carola-Bibiane Schönlieb.

As discussed in the previous chapter, discrete gradient methods yield unconditionally stable optimisation schemes when applied to the gradient flow (1.12). While these methods are well understood in the setting of geometric numerical integration, only in recent years have they been considered as optimisation schemes, and thus the analysis is lacking in this context. In this chapter, we seek to lay the foundations for discrete gradient methods for smooth optimisation, providing a comprehensive analysis of the well-posedness of the discrete gradient equation (2.8), optimal choices of time steps τ_k, convergence rates for different classes of functions, and guarantees of convergence to a unique limit.

We thus consider the unconstrained optimisation problem

$$\min_{x \in \mathbb{R}^n} F(x), \tag{3.1}$$

where the function $F : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

3.1.1 Contributions and outline

While discrete gradient methods have existed in geometric integration since the 1980s, only recently have they been studied in the context of optimisation, leaving significant gaps in our understanding of these schemes. In this chapter, we resolve fundamental questions about the discrete gradient methods, including their well-posedness, efficiency, and optimal tuning.

In Section 3.2 we define discrete gradients and introduce the four discrete gradient methods considered in this thesis. In Section 3.3, we prove that the discrete gradient equation (the update formula) (2.8) is well-posed, meaning that for any time step $\tau_k > 0$ and $x^k \in \mathbb{R}^n$, a solution x^{k+1} exists, under mild assumptions on F. Using the Brouwer fixed point theorem, this is the first existence result for the discrete gradient equation without a bound on the time step. In Section 3.4, we propose an efficient and stable method for solving the discrete gradient equation and prove convergence guarantees.

In Section 3.5, we analyse the dependence of the iterates on the choice of time step, and obtain estimates for preferable time steps in the cases of L-smoothness and strong convexity. In Section 3.6, we establish convergence rates for convex functions with Lipschitz continuous gradients, and for functions that satisfy the Polyak–Łojasiewicz (PŁ) inequality [120]. In Section 3.7, we establish convergence guarantees for functions that satisfy the strong Kurdyka–Łojasiewicz inequality. In Section 3.9, we present numerical results for several test problems, and a numerical comparison of different numerical solvers for the discrete gradient equation (2.8).

We emphasise that the majority of these results hold for nonconvex functions.

3.2 Discrete gradient methods

In this chapter, we consider both deterministic schemes and stochastic schemes. For the stochastic schemes, there is a random distribution Ξ on S^{n-1} such that each iterate x^k depends on a descent direction d^k which is independently drawn from Ξ. We denote by ξ^k the joint distribution of $(d^i)_{i=1}^k$. We denote by F_{k+1} the expectation of $F(x^{k+1})$ conditioned on ξ^k,

$$F_{k+1} := \mathbb{E}_{\xi^k}[F(x^{k+1})]. \tag{3.2}$$

To unify notation for all the methods in this chapter, we will write F_{k+1} instead of $F(x^{k+1})$ for the deterministic methods as well.

We recall the three discrete gradient methods in Section 2.7, using the mean value, Gonzalez, and Itoh–Abe discrete gradients. In addition to these methods, we also propose a generalisation of the Itoh–Abe discrete gradient method, from hereon referred to as *randomised Itoh–Abe methods*. As the Itoh–Abe discrete gradient method is comparable to cyclic coordinate descent (CCD), the generalised method is comparable to randomised coordinate descent and random search methods.

Thus, we consider a sequence of independent, identically distributed directions $(d^k)_{k\in\mathbb{N}} \subset S^{n-1}$ drawn from a random distribution Ξ, and solve

$$x^{k+1} = x^k - \tau_k \frac{F(x^{k+1}) - F(x^k)}{\langle x^{k+1} - x^k, d^{k+1}\rangle} d^{k+1}, \tag{3.3}$$

This can be rewritten as solving

$$x^{k+1} \mapsto x^k - \tau_k \alpha_k d^{k+1}, \quad \text{where } \alpha_k \neq 0 \text{ solves} \quad \alpha_k = -\frac{F(x^k - \tau_k \alpha_k d^{k+1}) - F(x^k)}{\tau_k \alpha_k}, \tag{3.4}$$

where $x^{k+1} = x^k$ is considered a solution whenever $\langle \nabla F(x^k), d^{k+1}\rangle = 0$.

We also define the constant

$$\zeta := \min_{e\in S^{n-1}} \mathbb{E}_{d\sim\Xi}[\langle d, e\rangle^2], \tag{3.5}$$

and assume that Ξ is such that $\zeta > 0$. For example, for the uniform random distribution on both S^{n-1} and on the standard coordinates $(e^i)_{i=1}^n$, we have $\zeta = 1/n$. See [206, Table 4.1] for estimates of (3.5) for these cases and others.

This scheme is a generalisation of the Itoh–Abe discrete gradient method, in the sense that the methods are equivalent if $(d^k)_{k\in\mathbb{N}}$ cycle through the standard coordinates with the rule

$$d^k = e^{[(k-1) \bmod n]+1}, \quad k = 1, 2, \ldots$$

However, the computational effort of one iterate of the Itoh–Abe discrete gradient method is equal to n steps of the randomised method, so the efficiency of the methods should be judged accordingly. Furthermore, the dissipation properties (2.9) can be rewritten as

$$F(x^{k+1}) - F(x^k) = -\tau_k \left(\frac{F(x^{k+1}) - F(x^k)}{\|x^{k+1} - x^k\|}\right)^2 = -\frac{1}{\tau_k}\|x^{k+1} - x^k\|^2. \tag{3.6}$$

Consequently, the dissipative structure of the Itoh–Abe methods is well-defined in a derivative-free setting.

The motivation for introducing this randomised extension of the Itoh–Abe method is, first, to tie discrete gradient methods in with other optimisation methods such as randomised coordinate descent [89, 182, 221] and random pursuit [160, 206], and, second, because this method extends to the nonsmooth, nonconvex setting, as we will show in Chapter 4.

3.3 Existence of solutions to the discrete gradient steps

In this section, we prove that the discrete gradient equation

$$y = x - \tau \overline{\nabla} F(x,y). \tag{3.7}$$

admits a solution y, for all time steps $\tau > 0$ and points $x \in \mathbb{R}^n$, under mild assumptions on F and $\overline{\nabla} F$. The result applies to the three discrete gradients considered in this thesis, and we expect that it also covers a vast number of other discrete gradients. These results do not require convexity of F.

To the authors' knowledge, the following result is the first without a restriction on time steps. Norton and Quispel [164] provided an existence and uniqueness result for small time steps for a large class of discrete gradients, via the Banach fixed point theorem. Furthermore, the existence of a solution for the Gonzalez discrete gradient is established for sufficiently small time steps via the implicit function theorem in [207, Theorem 8.5.4].

Throughout this section, we consider a function operator $\overline{\nabla} : C^1(\mathbb{R}^n) \to C(\mathbb{R}^n \times \mathbb{R}^n; \mathbb{R}^n)$, which maps a function $F \in C^1(\mathbb{R}^n)$ to the discrete gradient $\overline{\nabla} F$. For a set $K \subset \mathbb{R}^n$ and $\delta > 0$, we define the δ-thickening, $K_\delta = \{x \in \mathbb{R}^n \, : \, \operatorname{dist}(K,x) \leq \delta\}$, where $\operatorname{dist}(K,x) := \inf_{y \in K} \|x - y\|$.

We make two assumptions for the discrete gradient operator $\overline{\nabla}$, namely that boundedness of the gradient implies boundedness of the discrete gradient, and that if two functions coincide on an open set, their discrete gradients also coincide.

Assumption 3.1. *There is a constant C_n that depends on the discrete gradient but is independent of F, and a continuous, nondecreasing function $\delta : [0,\infty] \to [0,\infty]$, where $\delta(0) = 0$, $\delta(r) < \infty$ for all $r < \infty$, and $\delta(\infty) := \lim_{r \to \infty} \delta(r)$, such that the following holds.*

For any $F \in C^1(\mathbb{R}^n)$ and any convex set $K \subset \mathbb{R}^n$ with nonempty interior, the two following properties are satisfied.

(i) If $\|\nabla F(x)\| \leq L$ for all $x \in K_{\delta(\operatorname{diam}(K))}$, then $\|\overline{\nabla} F(x,y)\| \leq C_n L$ for all $x,y \in K$.

(ii) If G is another continuously differentiable function such that $F(x) = G(x)$ for all $x \in K_{\delta(\operatorname{diam}(K))}$, then $\overline{\nabla} F(x,y) = \overline{\nabla} G(x,y)$ for all $x,y \in K$.

The following result shows that the three discrete gradients considered satisfy the above assumption.

Lemma 3.2. *The three discrete gradients satisfy Assumption 3.1 with the following constants.*

1. *For the Gonzalez discrete gradient, $C_n = \sqrt{2}$ and $\delta \equiv 0$.*

2. *For the mean value discrete gradient, $C_n = 1$ and $\delta \equiv 0$.*

3. *For the Itoh–Abe discrete gradient, $C_n = \sqrt{n}$ and $\delta(r) = r$.*

Proof. Part 1. We first consider the Gonzalez discrete gradient. Denote by d the unit vector $(y-x)/\|y-x\|$. There is a vector $d^\perp$ such that $\langle d, d^\perp\rangle = 0$, $\|d^\perp\| = 1$, and

$$\overline{\nabla}F(x,y) = \left\langle \nabla F\left(\frac{x+y}{2}\right), d^\perp \right\rangle d^\perp + \frac{F(y)-F(x)}{\|y-x\|}d.$$

By the mean value theorem, there is $z \in [x,y]$ such that $F(y) - F(x) = \langle \nabla F(z), y-x\rangle$. Therefore, we obtain

$$\overline{\nabla}F(x,y) = \left\langle \nabla F\left(\frac{x+y}{2}\right), d^\perp \right\rangle d^\perp + \langle \nabla F(z), d\rangle d. \tag{3.8}$$

From this, we derive

$$\|\overline{\nabla}F(x,y)\|^2 \leq \left\|\nabla F\left(\frac{x+y}{2}\right)\right\|^2 + \left\|\nabla F(z)\right\|^2.$$

This implies that property *(i)* holds with $C_n = \sqrt{2}$ and $\delta \equiv 0$. To show property *(ii)*, it is sufficient to note that since K is convex and has nonempty interior, then $\nabla G\big((x+y)/2\big) = \nabla F\big((x+y)/2\big)$.

Part 2. Next we consider the mean value discrete gradient. It is clear that property *(i)* holds with $C_n = 1$ and $\delta \equiv 0$. Property *(ii)* is immediate from convexity of K.

Part 3. For the Itoh–Abe discrete gradient, we set $\delta(r) = r$. By applying the mean value theorem to

$$[\overline{\nabla}F(x,y)]_i = \frac{F(y_1,\dots,y_i,x_{i+1},\dots,x_n) - F(y_1,\dots,y_{i-1},x_i,\dots,x_n)}{y_i - x_i}, \tag{3.9}$$

we derive that $(\overline{\nabla}F(x,y))_i = [\nabla F(z^i)]_i$, where $z^i = [y_1,\dots,y_{i-1},c_i,x_{i+1},\dots,x_n]^T$ for some $c_i \in [x_i, y_i]$. Furthermore, we have $\|z^i - x\| \leq \|y-x\|$, so $z \in K_{\mathrm{diam}(K)}$. This implies that property *(i)* holds with $C_n = \sqrt{n}$. Property *(ii)* is immediate. □

The existence proof is based on the Brouwer fixed point theorem [31], which we state here.

Proposition 3.3 (Brouwer fixed point theorem)**.** *Let $K \subset \mathbb{R}^n$ be a convex, compact set and $g : K \to K$ a continuous function. Then g has a fixed point in K.*

We proceed to state the existence theorem.

Theorem 3.4 (Discrete gradient existence theorem)**.** *Suppose F is continously differentiable and that $\overline{\nabla}$ satisfies Assumption 3.1. Then there exists a solution y to* (3.7) *for any $\tau > 0$ and $x \in \mathbb{R}^n$, if F satisfies* ***either*** *of the following properties.*

(i) The gradient of F is uniformly bounded.

(ii) F is coercive.

(iii) Both F and the gradient of F are uniformly bounded on $\operatorname{co}(\{y \,:\, F(y) \leq F(x)\})$ *(the bounds may depend on x), and $\delta \equiv 0$ in Assumption 3.1.*

Proof. Part *(i)*. We define the function $g(y) = x - \tau\overline{\nabla}F(x,y)$, and want to show that it has a fixed point, $y = g(y)$. There is $L > 0$ such that $\|\nabla F(y)\| \leq L$ for all $y \in \mathbb{R}^n$. Therefore, by Assumption 3.1, $\|\overline{\nabla}F(x,y)\| \leq C_n L$ for all $y \in \mathbb{R}^n$. This implies that $g(y) \in \overline{B}_{\tau C_n L}(x)$ for all $y \in \mathbb{R}^n$. Specifically, g maps $\overline{B}_{\tau C_n L}(x)$ into itself. As g is continuous, it follows from the Brouwer fixed point theorem that there exists a point $y \in \overline{B}_{\tau C_n L}(x)$ such that $g(y) = y$, and we are done.

Part *(ii)*. Let $\sigma > 0$, $K = \operatorname{co}(\{y \,:\, F(y) \leq F(x)\})$, and write $\delta = \delta(\operatorname{diam}(K))$. Since F is coercive, K_δ and $K_{\delta+\sigma}$ are bounded. By standard arguments [161, Corollary 2.5], there exists a cutoff function $\varphi \in C_c^\infty(\mathbb{R}^n; [0,1])$ such that

$$\varphi(y) = \begin{cases} 1 & \text{if } y \in K_\delta, \\ 0 & \text{if } y \notin K_{\delta+\sigma}. \end{cases}$$

We define $G : \mathbb{R}^n \to \mathbb{R}$ by $G(y) := \varphi(y)\big(F(y) - F(x)\big) + F(x)$. G is continuously differentiable and $\operatorname{supp}(\nabla G) \subset K_{\delta+\sigma}$. Therefore, G has uniformly bounded gradient, so by part *(i)* there is a y such that

$$y = x - \tau\overline{\nabla}G(x,y).$$

By (2.9), $G(y) < G(x)$ which implies that $y \in K_\delta$, so $G(y) = F(y)$. Furthermore, since $G(x) = F(x)$, we deduce that $F(y) < F(x)$, so $y \in K$. Lastly, since F and G coincide on K_δ, and x and y both belong to K, it follows from Assumption 3.1 *(ii)* that $\overline{\nabla}F(x,y) = \overline{\nabla}G(x,y)$. Hence a solution $y = x - \tau\overline{\nabla}F(x,y)$ exists.

Part *(iii)*. Set $K = \operatorname{co}\Big(\{y \,:\, F(y) \leq F(x)\}\Big)$ and $M = \sup_{y \in K} F(y)$. Furthermore let $\varepsilon > 0$ and set $L = \sup\{\|\nabla F(y)\| \,:\, F(y) \leq M + \varepsilon\}$ and $F = \{y \,:\, F(y) \geq M + \varepsilon\}$. The mean value theorem [162, Equation A.55] and the boundedness of ∇F imply that for all $y \in K$ and $z \in F$, there is $\lambda \in (0,1)$ such that

$$\varepsilon \leq |F(y) - F(z)| = |\langle \nabla F(\lambda y + (1-\lambda)z), y - z\rangle| \leq L\|y - z\|.$$

Therefore, for all $y \in K$ and $z \in F$, $\|y-z\| \geq \varepsilon/L$. By Lemma A.1, there exists a cutoff function $\varphi \in C^\infty(\mathbb{R}^n;[0,1])$ with uniformly bounded gradient, such that

$$\varphi(y) = \begin{cases} 1 & \text{if } y \in K, \\ 0 & \text{if } y \in F. \end{cases}$$

Consider $G : \mathbb{R}^n \to \mathbb{R}$ defined as in the previous case. The gradient of G is uniformly bounded, so there is a fixed point y such that $y = x - \tau\overline{\nabla}G(x,y)$. By the same arguments as in case *(ii)*, $\overline{\nabla}F(x,y) = \overline{\nabla}G(x,y)$, which implies that y solves $y = x - \tau\overline{\nabla}F(x,y)$. □

The third case in Theorem 3.4 covers optimisation problems where F is not coercive. This includes the cases of linear systems with nonempty kernel and logistic regression problems [129] without regularisation.

While the above theorem also covers the Itoh–Abe methods, there is a much simpler existence result in this case, given in Chapter 4. This requires only continuity of the objective function, rather than differentiability.

3.4 Solving the discrete gradient equation

In the previous section, we proved that the discrete gradient equation (3.7)

$$y = x - \tau\overline{\nabla}F(x,y),$$

admits a solution y for all $\tau > 0$ and $x \in \mathbb{R}^n$. In what follows, we discuss how to approximate a solution to (3.7) when no closed-form expression exists, using fixed point iterations. We do not consider the Itoh–Abe discrete gradient, which simply involve solving successive scalar equations.

Norton and Quispel [164] showed that for a given $x \in \mathbb{R}^n$ and sufficiently small time steps, there exists a unique solution to (3.7) that can be approximated by the fixed point iterations

$$y^{k+1} = T_\tau(y^k), \qquad \text{where} \qquad T_\tau(y) := x - \tau\overline{\nabla}F(x,y). \tag{3.10}$$

That is, the iterates converge to the fixed point $y^* = T_\tau(y^*)$, i.e. a solution to (3.7). Their analysis assumes that the time step τ is less than $1/(10L_{\mathrm{DG}})$, where L_{DG} is the Lipschitz constant for a given x of the mapping $y \mapsto \overline{\nabla}F(x,y)$.

However, for optimisation, we are interested in larger time steps (for L-smooth functions, the optimal time steps are typically around $2/L$—see Section 3.6), while it is not so important to have uniqueness of solutions to (3.7). Furthermore, as Theorem 3.4 ensures the existence of a solution for arbitrarily large time steps, we seek a constructive method for locating such solutions. We therefore propose the following relaxation of the fixed point updates. For $\theta \in (0,1]$, update

$$y^{k+1} = (1-\theta)y^k + \theta T_\tau(y^k). \tag{3.11}$$

For $\theta = 1$, this reduces to (3.10). In the remainder of this section, we will prove convergence guarantees of (3.11) for all time steps. In Section 3.9, we demonstrate its numerical efficiency.

In the following, we assume that the discrete gradient inherits smoothness and strong convexity properties from the gradient. As with the previous section, we here consider the discrete gradient as a function operator $\overline{\nabla} : C^1(\mathbb{R}^n) \to C(\mathbb{R}^n \times \mathbb{R}^n; \mathbb{R}^n)$.

Assumption 3.5. *There is $\lambda_L, \lambda_\mu > 0$, such that the discrete gradient operator $\overline{\nabla}$ satisfies:*

(i) (Smoothness) *If F is L-smooth, then for all $x \in \mathbb{R}^n$, $y \mapsto \overline{\nabla}F(x,y)$ is $\lambda_L L$-smooth.*

(ii) (Monotonicity) *If F is μ-convex, then for all $x,y,z \in \mathbb{R}^n$, we have*

$$\langle \overline{\nabla}F(x,y) - \overline{\nabla}F(x,z), y-z\rangle \geq \lambda_\mu \mu \|y-z\|^2.$$

We write $L_{DG} := \lambda_L L$ and $\mu_{DG} := \lambda_\mu \mu$.

Remark 3.6. *It always holds that $\mu_{DG} \leq L_{DG}$.*

It is trivial to show these properties for the mean value discrete gradient.

Proposition 3.7. *The mean value discrete gradient satisfies Assumption 3.5 with $L_{DG} = L/2$ and $\mu_{DG} = \mu/2$.*

Proof. To show that the first property holds, we write

$$\begin{aligned}\|\overline{\nabla}F(x,y) - \overline{\nabla}F(x,z)\| &\leq \int_0^1 \|\nabla F(sy+(1-s)x) - \nabla F(sz+(1-s)x)\| \, \mathrm{d}s \\ &\leq L\|y-z\| \int_0^1 s \, \mathrm{d}s = \frac{L}{2}\|y-z\|.\end{aligned}$$

Similarly, to show the second property, we write

$$\begin{aligned}\langle \overline{\nabla}F(x,y) - \overline{\nabla}F(x,z), y-z\rangle &= \int_0^1 \frac{1}{s} \big\langle \nabla F(sy+(1-s)x) - \nabla F(sz+(1-s)x), sy-sz \big\rangle \, \mathrm{d}s \\ &\geq \mu\|y-z\|^2 \int_0^1 s \, \mathrm{d}s = \frac{\mu}{2}\|y-z\|^2.\end{aligned}$$

□

Remark 3.8. *We were unable to ascertain whether or not the properties hold for the Gonzalez discrete gradient. However, we observe in practice that the scheme converges in this case too.*

The following result demonstrates that for convex objective functions, the scheme (3.11) converges to a fixed point $y^* = T_\tau(y^*)$ for arbitrary time steps τ.

Theorem 3.9. *If F is L-smooth and $\overline{\nabla}$ satisfies Assumption 3.5, then for any $x \in \mathbb{R}^n$ the iterates $(y^k)_{k\in\mathbb{N}}$ defined by* (3.11) *converge linearly to a fixed point $y^* = T_\tau(y^*)$ if either of the following cases hold.*

(i) $\tau < 1/L_{DG}$.

(ii) *F is μ-convex and* $\theta \in (0, \min\{1, \frac{2+2\tau\mu_{DG}}{1+\tau^2 L_{DG}^2+2\tau\mu_{DG}}\})$.

Proof. Case (i). We write

$$\begin{aligned}\|y^{k+1}-y^k\| &= \|(1-\theta)(y^k-y^{k-1})+\tau\theta(\overline{\nabla}F(x,y^{k-1})-\overline{\nabla}F(x,y^k)\| \\ &\leq \big(1-(1-\tau L_{\mathrm{DG}})\theta\big)\,\|y^k-y^{k-1}\|.\end{aligned}$$

This converges whenever $1-(1-\tau L_{\mathrm{DG}})\theta < 1$, i.e. when $\tau < 1/L_{\mathrm{DG}}$.

Case (ii). In a similar fashion, we write

$$\begin{aligned}\|y^{k+1}-y^k\|^2 &= \|(1-\theta)(y^k-y^{k-1})+\tau\theta(\overline{\nabla}F(x,y^{k-1})-\overline{\nabla}F(x,y^k)\|^2 \\ &= (1-\theta)^2\|y^k-y^{k-1}\|^2+\tau^2\theta^2\|\overline{\nabla}F(x,y^{k-1})-\overline{\nabla}F(x,y^k)\|^2 \\ &\qquad -2\tau(1-\theta)\theta\langle y^k-y^{k-1},\overline{\nabla}F(x,y^k)-\overline{\nabla}F(x,y^{k-1})\rangle \\ &\leq \underbrace{\Big((1-\theta)^2+\tau^2\theta^2L_{\mathrm{DG}}^2-2\tau(1-\theta)\theta\mu_{\mathrm{DG}}\Big)}_{\omega(\theta)}\|y^k-y^{k-1}\|^2.\end{aligned}$$

One can check that the coefficient $\omega(\theta)$ is less than 1 provided θ belongs to the interval stated in the theorem. This concludes the proof. □

Remark 3.10. *In the second case of the above theorem, the coefficient $\omega(\theta)$ is minimised for*

$$\theta^* = \frac{1+\tau\mu_{DG}}{1+\tau^2L_{DG}^2+2\tau\mu_{DG}} < 1, \tag{3.12}$$

which yields the linear convergence rate

$$\|y^{k+1}-y^k\|^2 \leq \frac{\tau^2(L_{DG}^2-\mu_{DG}^2)}{(1+\tau\mu_{DG})^2+\tau^2(L_{DG}^2-\mu_{DG}^2)}\|y^k-y^{k-1}\|^2.$$

We note from this that the scheme converges faster for smaller time steps and for objective functions with smaller condition numbers $L/\mu \approx L_{DG}/\mu_{DG} =: \kappa_{DG}$. Furthermore, if $\tau = 1/(aL_{DG})$ for some $a \geq 1$, where a typical choice is $a = 1$, then we obtain

$$\theta^* = \frac{1 + \frac{1}{a\kappa_{DG}}}{1 + \frac{1}{a^2} + \frac{2}{a\kappa_{DG}}} \geq \frac{a^2}{1+a^2}, \quad \omega(\theta^*) = \frac{1 - \frac{1}{\kappa_{DG}^2}}{a^2 + \frac{2a}{\kappa_{DG}} + 1} \leq \frac{1}{a^2+1}.$$

This shows that the fixed point scheme (3.11) *is robust to ill-conditioned problems, both with regards to appropriate choices of θ and the rate of convergence.*

In Section 3.9.6, we compare the efficiency of the above scheme for different θ and of the built-in solver `scipy.optimize.fsolve` in Python.

3.5 Analysis of time steps for discrete gradient methods

In this section, we study the implicit dependence of $x^{k+1}(\tau)$ on the choice of time step τ. We concentrate on the mean value and Itoh–Abe discrete gradient methods, establishing a uniqueness result for the update assuming convexity, as well as bounds on optimal time steps with respect to the decrease in F, for L-smooth, convex functions as well as strongly convex functions.

3.5.1 Uniqueness for convex objectives

Lemma 3.11. *If F is convex, then the solution y to the discrete gradient equation* (3.7) *is unique for the mean value discrete gradient and the Itoh–Abe discrete gradient.*

Proof. We first consider the mean value discrete gradient. Suppose there are two solutions $y^1, y^2 \in \mathbb{R}^n$ to (3.7), i.e.

$$y^i = x - \tau\overline{\nabla}F(x,y^i), \quad i = 1,2.$$

Then

$$\begin{aligned}\|y^1 - y^2\|^2 &= \tau\left\langle \frac{x-y^2}{\tau} - \frac{x-y^1}{\tau}, y^1 - y^2 \right\rangle \\ &= \tau\langle \overline{\nabla}F(x,y^2) - \overline{\nabla}F(x,y^1), y^1 - y^2\rangle \leq 0,\end{aligned}$$

where the last inequality follows from Proposition 3.7.

To show uniqueness of the Itoh–Abe discrete gradient method, we note that the Itoh–Abe update consists of a succession of scalar updates, and that for scalar problems all discrete gradients are the same. Hence uniqueness is inherited from uniqueness of the mean value discrete gradient. □

3.5.2 Implicit dependence on the time step for mean value discrete gradient methods

In this subsection, we consider the mean value discrete gradient and study the dependence of the update (3.7) on the choice of time step for convex functions. While we assume in the proofs that F is C^2-smooth, the statements can be generalised to C^1-smooth functions, as we can consider a sequence of C^2-smooth functions $F_k : \mathbb{R}^n \to \mathbb{R}$ such that $\|F_k - F\|_1 \to 0$ and $\|\nabla F_k - \nabla F\|_1 \to 0$ (which exist by [84, Section 5.2, Theorem 3]), and pass to the limit for the discrete gradient equation.

In the previous subsection, we showed that the update (3.7) is unique for convex F. It follows that for a given x, we can consider the unique mapping $\tau \mapsto y_\tau$ implicitly defined by

$$y_\tau = x - \tau \overline{\nabla} F(x, y_\tau).$$

It is straightforward to show that it is a continuous path, by arguing similarly to the proof of Proposition 3.7. We furthermore want to show that it is differentiable. To do so, we can use the implicit function theorem Proposition 2.3.

Define the function

$$G : \mathbb{R}^n \times (0,\infty) \mapsto \mathbb{R}^n, \quad G(y,\tau) = y - x + \tau \overline{\nabla} F(x,y), \tag{3.13}$$

so that y_τ is the unique solution to $G(y_\tau, \tau) = 0$. Assuming that F is C^2-smooth, then G is C^2-smooth, and the gradients are given by

$$\nabla_y G(y,\tau) = I + \tau \int_0^1 s \nabla^2 F((1-s)x + sy)\,\mathrm{d}s, \qquad \nabla_\tau G(y,\tau) = \overline{\nabla} F(x,y).$$

Since F is convex, the Hessian $\nabla^2 F$ is positive-definite, so $\nabla_y G(y,\tau)$ is invertible for all y, τ. By the implicit function theorem Proposition 2.3, we conclude with the following.

Proposition 3.12. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be convex and C^2-smooth, and let G be defined by* (3.13). *Then mapping $\tau \mapsto y_\tau$ is C^1-smooth, and its gradient is given by*

$$Dy_\tau = -\left(I + \tau \int_0^1 s \nabla^2 F((1-s)x + sy_\tau)\,\mathrm{d}s \right)^{-1} \overline{\nabla} F(x, y_\tau). \tag{3.14}$$

For notational brevity, we write $H_\tau := \int_0^1 s\nabla^2 F((1-s)x + sy_\tau)\,\mathrm{d}s$.

From this, we can derive a number of properties of the dependence on the time step. First we show that the distance from x to y_τ strictly increases with τ.

Lemma 3.13. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a C^1-smooth and convex function. If $\tau_2 > \tau_1$, then $\|y_{\tau_2} - x\| > \|y_{\tau_1} - x\|$.*

Proof. We suppose F is C^2-smooth and calculate

$$\begin{aligned} D_\tau \frac{1}{2}\|y_\tau - x\|^2 = \langle Dy_\tau, y_\tau - x\rangle &= -\left\langle (I + \tau H_\tau)^{-1}\overline{\nabla}F(x,y_\tau), y_\tau - x\right\rangle \\ &= \frac{1}{\tau}\left\langle (I+\tau H_\tau)^{-1}(y_\tau - x), y_\tau - x\right\rangle > 0, \end{aligned}$$

where the last inequality follows from positive-definiteness of $(I + \tau H_\tau)^{-1}$. The result follows for C^2-smooth functions F, and as explained above, we can apply an approximation argument to extend the result to C^1-smooth functions. □

Next we analyse the dependence of $F(y_\tau)$ on τ for L-smooth and μ-convex functions. We make use of the following properties.

Proposition 3.14. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be C^2-smooth, L-smooth and μ-convex, for $\mu \geq 0$. Denote by A_τ the matrix $(I + \tau H_\tau)^{-1}$, and by κ_τ the condition number of A_τ. Then F and A_τ satisfy the following properties.*

(i) ***(Norms:)*** $\|A_\tau\| \leq 1/(1+\tau\mu/2)$ *and* $\|A_\tau^{-1}\| \leq 1 + \tau L/2$.

(ii) ***(Descent lemma:)*** $\langle \nabla F(y), A_\tau(x-y)\rangle \geq F(x) - F(y) - \frac{L}{2}\sqrt{\kappa_\tau}\langle x-y, A_\tau(x-y)\rangle$.

(iii) ***(Convexity:)*** $\langle \nabla F(y), A_\tau(x-y)\rangle \leq F(y) - F(x) - \frac{\mu}{2\|A_\tau\|}\langle x-y, A_\tau(x-y)\rangle$.

Proof. Case *(i)*. Consider the inverse of A_τ, $I + \tau H_\tau$. We know that $\|A_\tau\| = 1/\sigma_1$, where σ_1 is the smallest singular value of $I + \tau H_\tau$. Since F is μ-convex, it follows from Proposition 3.7 that $\sigma_1 \geq 1 + \tau\mu/2$, which yields the first bound. Similarly, using L-smoothness of F, it is straightforward to derive the second bound.

Case *(ii)*. Consider the inner product $\langle x,y\rangle_{A_\tau} := \langle x, A_\tau y\rangle$ with its associated norm $\|\cdot\|_{A_\tau}$. One can show that if F is L-smooth with respect to the norm $\|\cdot\|$, then F is $L\sqrt{\kappa_\tau}$-smooth with respect to $\|\cdot\|_{A_\tau}$. Thus we obtain the desired inequality by applying the regular descent lemma Proposition 2.39 with respect to the inner product $\langle\cdot,\cdot\rangle_{A_\tau}$.

Case *(iii)*. By μ-convexity, for all $x, y \in \mathbb{R}^n$ and $s \in (0,1)$, we have

$$\begin{aligned} F(sx + (1-s)y) &\leq sF(x) + (1-s)F(y) - \frac{\mu}{2}s(1-s)\langle x-y, x-y\rangle \\ &\leq sF(x) + (1-s)F(y) - \frac{\mu}{2\|A_\tau\|}s(1-s)\langle x-y, A_\tau(x-y)\rangle \end{aligned}$$

Hence, F is $\mu/\|A_\tau\|$-convex with respect to the inner product $\langle\cdot,\cdot\rangle_{A_\tau}$, and the result follows. □

We proceed to show that if F is L-smooth, then for $\tau \in (0, \frac{2}{L}\sqrt{2\kappa_F/(\kappa_F+1)})$, the function value $F(y_\tau)$ is decreasing with respect to τ. Here κ_F denotes the conditioning number of F, L/μ.

Lemma 3.15. *If $\tau_1 < \tau_2 \leq \frac{2}{L}\sqrt{2\kappa_F/(\kappa_F+1)}$, then $F(y_{\tau_1}) > F(y_{\tau_2})$.*

Proof. We calculate

$$\begin{aligned}
\mathrm{d}_\tau F(y_\tau) &= \langle \nabla F(y_\tau), \mathrm{d}_\tau y_\tau\rangle = -\left\langle \nabla F(y_\tau), (I+\tau H_\tau)^{-1}\overline{\nabla} F(x,y_\tau)\right\rangle \\
&= -\frac{1}{\tau}\left\langle \nabla F(y_\tau), (I+\tau H_\tau)^{-1}(x-y_\tau)\right\rangle \\
&\leq \frac{1}{\tau}\left(F(y_\tau) - F(x) + \frac{L}{2}\sqrt{\kappa_{A_\tau}}\langle x-y_\tau, A_\tau(x-y_\tau)\rangle\right) \\
&= \frac{1}{\tau}\left(-\frac{1}{\tau}\|x-y_\tau\|^2 + \frac{L}{2}\sqrt{\kappa_{A_\tau}}\langle x-y_\tau, A_\tau(x-y_\tau)\rangle\right) \leq \frac{1}{\tau}\left(\frac{L}{2}\sqrt{\kappa_{A_\tau}} - \frac{1}{\tau}\right)\|x-y_\tau\|^2.
\end{aligned}$$

This is strictly negative if $\tau < 2/(L\sqrt{\kappa_{A_\tau}})$. We have

$$\kappa_{A_\tau} = \frac{1+\tau\frac{L}{2}}{1+\tau\frac{\mu}{2}} \leq \kappa_{A_\tau} \leq \frac{2\kappa_F}{\kappa_F+1},$$

since the condition number strictly increases with τ, and $\tau < 2/L$. This concludes the proof. □

Finally, we show that if F is μ-convex for $\mu > 0$, then for

$$\tau > \frac{\kappa_F - 1 + \sqrt{(\kappa_F-1)^2+4}}{\mu}, \tag{3.15}$$

the function value $F(y_\tau)$ is increasing with respect to τ.

Lemma 3.16. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-convex for $\mu > 0$. If $\tau_2 > \tau_1$ and τ_1 satisfies* (3.15)*, then $F(y_{\tau_2}) > F(y_{\tau_1})$.*

Proof. We suppose F is C^2-smooth and that τ satisfies (3.15), and calculate

$$\begin{aligned}
D_\tau F(y_\tau) &= -\left\langle \nabla F(y_\tau), (I+\tau H_\tau)^{-1}\overline{\nabla}F(x,y_\tau)\right\rangle = \frac{1}{\tau}\left\langle \nabla F(y_\tau), (I+\tau H_\tau)^{-1}(y_\tau - x)\right\rangle \\
&\geq \frac{1}{\tau}\left(F(y_\tau) - F(x) + \frac{\mu}{2\|A_\tau\|}\langle x - y_\tau, A_\tau(x-y_\tau)\rangle\right) \\
&= \frac{1}{\tau}\left(\left\langle x - y_\tau, \left(\frac{\mu}{2\|A_\tau\|}A_\tau - \frac{1}{\tau}I\right)(x-y_\tau)\right\rangle\right).
\end{aligned}$$

The operator $\frac{\mu}{2\|A\|}A - \frac{1}{\tau}I$ is positive-definite, provided $\tau > \frac{2}{\mu}\|A_\tau\|\|A_\tau^{-1}\|$. By solving for τ, we can show that this follows from (3.15). □

The above analysis shows that the updates of the discrete gradient method behave as one would expect, with respect to time step, and furthermore provides us with a sense of optimal time step choices. In the following subsection, we give a similar analysis for the Itoh–Abe methods, but with sharper bounds on the optimal time steps. Furthermore, one may compare the bounds derived in this section, with those derived for the convergence rate analysis, e.g. in Lemma 3.21.

3.5.3 Implicit dependence on the time step for Itoh–Abe methods

For the remainder of the section, we restrict our focus to Itoh–Abe methods. We fix a starting point x, direction $d \in S^{n-1}$ and time step τ, and study the solution y to

$$y = x - \alpha d, \qquad \text{where } \alpha \neq 0 \text{ solves} \qquad \alpha = -\tau\frac{F(x-\alpha d) - F(x)}{\alpha}. \tag{3.16}$$

By the analysis in Section 3.3, there exists a solution y for all $\tau > 0$. For convenience and to exclude the case $y = x$, we assume $\langle \nabla F(x), d\rangle > 0$. For notational brevity, we rewrite the optimisation problem in terms of a scalar function f, i.e. solve

$$\frac{f(\alpha)}{\alpha^2} = -\frac{1}{\tau}, \qquad \text{where } f(\alpha) := F(x-\alpha d) - F(x). \tag{3.17}$$

For optimisation schemes with a time step τ, it is common to assume that the distance between x and y increases with the time step. For explicit schemes, this naturally holds, and in the previous subsection, we showed that it holds for the Itoh–Abe discrete gradient method on convex functions. However, for general functions this is not necessarily the case, as the following example shows.

Example 3.17. *Define $F(x) := -x^3$ and $x = 0$. For all $\tau > 0$,* (3.16) *is solved by*

$$y = \frac{1}{\tau}.$$

Then, as $\tau \to 0$, we have $y \to \infty$, and as $\tau \to \infty$, we have $y \to x$.

The above example illustrates that for nonconvex functions, decreasing the time step might lead to a larger step $y \leftarrow\!\shortmid x$ and vice versa.

The remainder of this section is devoted to deriving bounds on optimal time steps, with respect to the decrease in the objective function when the objective function is L-smooth or μ-convex. We first consider L-smooth functions, and show that any time step $\tau < 2/L$ is suboptimal. We recall the scalar function $f(\alpha) = F(x - \alpha d) - F(x)$. The following statement is the scalar version of Lemma 3.15, noting that in the scalar case, we can set $\kappa_{A_\tau} = 1$ and update the analysis in the proof.

Lemma 3.18. *If F is convex and L-smooth, then $\tau \mapsto f(\alpha(\tau))$ is strictly decreasing for $\tau \in (0, 2/L)$.*

We next show that for strongly convex functions, any time step $\tau > 2/\mu$ yields a suboptimal decrease. This is the scalar version of Lemma 3.16. We provide a separate proof with a sharper bound on the time step.

Lemma 3.19. *If F is μ-convex with $\mu > 0$, then $\tau \mapsto f(\alpha(\tau))$ is strictly increasing for $\tau > 2/\mu$.*

Proof. Let α solve (3.17) for $\tau > 2/\mu$. Fix $\lambda \in (2/(\tau\mu), 1)$, and plug in 0 and α for y and x respectively in Definition 2.17 *(ii)* to get, after rearranging,

$$f(\lambda\alpha) \leq \lambda f(\alpha) - \frac{\mu\lambda(1-\lambda)}{2}\alpha^2.$$

Plugging in (3.17) gives us

$$f(\lambda\alpha) \leq \left(\lambda + \frac{\tau\mu\lambda(1-\lambda)}{2}\right) f(\alpha).$$

We want to show that $f(\lambda\alpha) < f(\alpha)$, i.e. that $\lambda + \tau\mu\lambda(1-\lambda)/2 > 1$. By rearranging and solving the quadratic expression, we find that this is satisfied if $\lambda \in (2/(\tau\mu), 1)$. The result follows from convexity of f and Lemma 3.13. □

Remark 3.20. *This result also holds for strongly convex, non-differentiable functions.*

3.6 Convergence rate analysis

In this section we derive convergence rates for L-smooth, convex functions, μ-convex functions, and more generally functions that satisfy the Polyak–Łojasiewicz (PŁ) inequality. We follow the arguments in [15, 159] on convergence rates of coordinate descent.

We recall the notation in (3.2), $F_{k+1} := \mathbb{E}_{\xi^k} F(x^{k+1})$, where $F_{k+1} = F(x^{k+1})$ for deterministic methods. Estimates of the following form will be crucial to the analysis, for some descent constant $\beta > 0$.

$$\beta \left(F(x^k) - F_{k+1} \right) \geq \|\nabla F(x^k)\|^2 \tag{3.18}$$

We first provide this estimate for each of the four methods. We assume throughout that the time steps $(\tau_k)_{k\in\mathbb{N}}$ satisfy arbitrary bounds (2.10).

We consider coordinate-wise Lipschitz constants for the gradient of F as well as a directional Lipschitz constant. For $i = 1, \dots, n$, we suppose $[\nabla F]_i : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $\overline{L}_i \leq L$. We denote by $\overline{L}_{\text{sum}}$ the ℓ^2-norm of the coordinate-wise Lipschitz constants, $\overline{L}_{\text{sum}} = \sqrt{\sum_{i=1}^n \overline{L}_i^2} \in [L, \sqrt{n}L]$.

Furthermore, for a direction $d \in S^{n-1}$, we consider the Lipschitz constant $L_d \leq L$, such that for all $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, we have

$$|\langle \nabla F(x + \alpha d), d\rangle - \langle \nabla F(x), d\rangle| \leq L_d |\alpha|.$$

For the Itoh–Abe discrete gradient method or when Ξ only draws from the standard coordinates, we write L_i instead of L_{e^i}. We define $L_{\max} \leq L$ to be the supremum of L_d over all d in the support of the probability density function of Ξ. That is, $L_{\max} \geq L_d$ for all $d \sim \Xi$. In the case where Ξ draws from a restricted set, such as the standard coordinates, $L_{\max}$ can be notably smaller than L. In this setting, we can refine the L-smoothness property in Proposition 2.39 *(i)* to

$$F(x + \alpha d) - F(x) \leq \alpha \langle \nabla F(x), d\rangle + \frac{L_d}{2}\alpha^2 \leq \alpha \langle \nabla F(x), d\rangle + \frac{L_{\max}}{2}\alpha^2, \tag{3.19}$$

for all $\alpha \in \mathbb{R}$ and d in the support of the density of Ξ [15, Lemma 3.2].

Lemma 3.21. *If F is L-smooth, then the three discrete gradient methods and the randomised Itoh–Abe method satisfy* (3.18) *with values for β given in Table 3.1.*

Table 3.1 Estimates of β, as well as optimal time steps τ^* and corresponding β^*. Recall ζ is defined in (3.5).

Discrete gradient method	β	τ^*	β^*
Gonzalez	$2\left(1/\tau_k + L^2\tau_k/2\right)$	$\sqrt{2}/L$	$2\sqrt{2}L$
Mean value	$2\left(1/\tau_k + L^2\tau_k/4\right)$	$2/L$	$2L$
Itoh–Abe	$2\left(1/\tau_k + \overline{L}^2_{\text{sum}}\tau_k\right)$	$1/\overline{L}_{\text{sum}}$	$4\overline{L}_{\text{sum}}$
Randomised Itoh–Abe	$\tau_k\left(1/\tau_k + L_{\max}/2\right)^2/\zeta$	$2/L_{\max}$	$2L_{\max}/\zeta$

Proof. Part 1. We consider the characterisation (3.8) of the Gonzalez discrete gradient to compute

$$\begin{aligned}\|\nabla F(x^k)\|^2 &= \langle \nabla F(x^k), d\rangle^2 + \langle \nabla F(x^k), d^\perp\rangle^2 \\ &\leq 2\Big(\|\overline{\nabla} F(x^k, x^{k+1})\|^2 + \langle \nabla F(x^k) - \nabla F(z), d\rangle^2 + \\ &\qquad\qquad \left\langle \nabla F(x^k) - \nabla F\left(\tfrac{x^k + x^{k+1}}{2}\right), d^\perp \right\rangle^2\Big) \\ &\leq 2\left(\|\overline{\nabla} F(x^k, x^{k+1})\|^2 + \langle \nabla F(x^k) - \nabla F(z), d\rangle^2 + \frac{1}{4}L^2\|x^k - x^{k+1}\|^2\right).\end{aligned}$$

Since $\langle \nabla F(z), d\rangle = (F(x^{k+1}) - F(x^k))/\|x^{k+1} - x^k\|$ and $d = \frac{x^{k+1} - x^k}{\|x^{k+1} - x^k\|}$, we have

$$\begin{aligned}\langle \nabla F(x^k) - \nabla F(z), d\rangle^2 &= \frac{1}{\|x^k - x^{k+1}\|^2}\left(\langle \nabla F(x^k), x^{k+1} - x^k\rangle - F(x^{k+1}) + F(x^k)\right)^2 \\ &\leq \frac{1}{4}L^2\|x^{k+1} - x^k\|^2,\end{aligned}$$

where the inequality follows from Proposition 2.39 *(i)*. Therefore,

$$\|\nabla F(x^k)\|^2 \leq 2\left(\frac{1}{\tau_k} + \frac{1}{2}L^2\tau_k\right)\left(F(x^k) - F_{k+1}\right),$$

where we have used the discrete gradient properties (2.9).

Part 2. We compute

$$\begin{aligned}
\|\nabla F(x^k)\|^2 &\leq 2\|\overline{\nabla} F(x^k,x^{k+1})\|^2 + 2\left\|\int_0^1 \nabla F(sx^k + (1-s)x^{k+1}) - \nabla F(x^k)\,\mathrm{d}s\right\|^2 \\
&\leq 2\|\overline{\nabla} F(x^k,x^{k+1})\|^2 + 2L^2\|x^k - x^{k+1}\|^2 \left(\int_0^1 s\,\mathrm{d}s\right)^2 \\
&= 2\left(\frac{1}{\tau_k} + \frac{1}{4}L^2\tau_k\right)\left(F(x^k) - F_{k+1}\right).
\end{aligned}$$

Part 3. We apply the mean value theorem like in (3.9) to obtain $\left(\overline{\nabla} F(x^k,x^{k+1})\right)_i = [\nabla F(y^i)]_i$, where $y^i = [x_1^{k+1},\dots,x_{i-1}^{k+1},c_i,x_{i+1}^k,\dots,x_n^k]^T$ for $c_i \in [x_i^k, x_i^{k+1}]$. This gives

$$\begin{aligned}
\|\nabla F(x^k)\|^2 = \sum_{i=1}^n |[\nabla F(x^k)]_i|^2 &\leq 2\sum_{i=1}^n \left(|[\nabla F(y^i)]_i|^2 + |[\nabla F(y^i)]_i - [\nabla F(x^k)]_i|^2\right) \\
&\leq 2\left(\|\overline{\nabla} F(x^k,x^{k+1})\|^2 + \overline{L}_{\mathrm{sum}}^2\|x^k - x^{k+1}\|^2\right) \\
&\leq 2\left(\frac{1}{\tau_k} + \overline{L}_{\mathrm{sum}}^2\tau_k\right)\left(F(x^k) - F_{k+1}\right).
\end{aligned}$$

Part 4. By (3.19), we have

$$\begin{aligned}
\langle \nabla F(x^k), x^k - x^{k+1}\rangle &\leq F(x^k) - F(x^{k+1}) + \frac{L_{\max}}{2}\|x^k - x^{k+1}\|^2 \\
&= \left(\frac{1}{\tau_k} + \frac{L_{\max}}{2}\right)\|x^k - x^{k+1}\|^2,
\end{aligned}$$

where the second equality follows from (3.6).

Furthermore, $\langle \nabla F(x^k), x^k - x^{k+1}\rangle = |\langle \nabla F(x^k), d^{k+1}\rangle|\|x^k - x^{k+1}\|$. From this, we derive

$$\langle \nabla F(x^k), d^{k+1}\rangle^2 \leq \left(\frac{1}{\tau_k} + \frac{L_{\max}}{2}\right)^2 \|x^k - x^{k+1}\|^2. \tag{3.20}$$

By the definition of ζ, we have

$$\mathbb{E}_{d^{k+1}\sim\Xi}\langle \nabla F(x^k), d^{k+1}\rangle^2 \geq \zeta\|\nabla F(x^k)\|^2. \tag{3.21}$$

Combining (3.20) and (3.21), we derive

$$\|\nabla F(x^k)\|^2 \leq \frac{\tau_k}{\zeta}\left(\frac{1}{\tau_k} + \frac{L_{\max}}{2}\right)^2\left(F(x^k) - F_{k+1}\right).$$

This concludes the proof. □

Remark 3.22. *Note that these estimates do not require convexity of F. Also note that they immediately result in convergence rates for the gradient as well, inherited from the rates of the objective function.*

3.6.1 Optimal time steps and estimates of descent constant

Lower values for β in (3.18) correspond to better convergence rates, as can be seen in Theorems 3.25 and 3.27. In what follows, we briefly discuss the time steps that yield minimal values of β, denoted by τ^* and β^* in Table 3.1.

For the Gonzalez and mean value discrete gradient methods, it is natural to compare rates to those of explicit gradient descent, which has the estimate $\beta^* = 2L$ [158]. Hence, the mean value discrete gradient method recovers the optimal rates of gradient descent, while the estimate for the Gonzalez discrete gradient is worse by a factor of $\sqrt{2}$.

For the Itoh–Abe discrete gradient method, we compare its rates to those obtained for CCD schemes in [221, Theorem 3] and [15, Lemma 3.3],

$$\beta^* = 8\sqrt{n}L,$$

where we have set their parameters $L_{\max}$ and $L_{\min}$ to $\sqrt{n}L$. Hence, the estimate for the Itoh–Abe discrete gradient method is stronger, being at most half that of CCD, even in the worst-case scenario $\overline{L}_{\text{sum}} = \sqrt{n}L$.

Remark 3.23. *Note however that we can improve the estimate for the CCD scheme to recover the same rate. See Appendix A.2.*

We give one motivating example for considering the parameter $\overline{L}_{\text{sum}}$.

Example 3.24. *Let F be a least squares problem $F(x) = \|Ax - f\|^2/2$. We then have*

$$\overline{L}_{\text{sum}} \leq \sqrt{\operatorname{rank}(A)}L. \tag{3.22}$$

Thus, for low-rank system where $\operatorname{rank}(A) \ll n$, the convergence speed of the Itoh–Abe discrete gradient method improves considerably.

To derive (3.22), one can show that $L = \|A^*A\|$ and $\overline{L}_{\text{sum}} = \|A^*A\|_F$. The bound then follows from the fact that $\|B\|_F \leq \sqrt{\operatorname{rank}(B)}\|B\|$ [109, Table 6.2] and that $\operatorname{rank}(A^*A) = \operatorname{rank}(A)$ [149, Statement 4.5.4].

We compare the rates for the randomised Itoh–Abe methods to randomised coordinate descent (RCD). Recall that when Ξ is the random uniform distribution on the coordinates $(e^i)_{i=1}^n$ or on the unit sphere S^{n-1}, we have $\zeta = 1/n$. This gives us $\beta^* = 2nL_{\max}$ for the randomised Itoh–Abe methods, which is the optimal bound for randomised coordinate descent [221, Equation 30].

3.6.2 Lipschitz continuous gradients

For the next result, we use the notation $R(x^0) = \operatorname{diam}\left\{x \in \mathbb{R}^n \,:\, F(x) \leq F(x^0)\right\}$. This is bounded, provided F is coercive.

Theorem 3.25. *Let F be an L-smooth, convex, coercive function. Then for all four methods, we have*

$$F_k - F^* \leq \frac{\beta R(x^0)^2}{k + 2\frac{\beta}{L}}.$$

where β is given in Table 3.1 and $F^* := \min_x F(x)$.

Proof. Let x^* be a minimizer of F. By respectively convexity, the Cauchy-Schwarz inequality, and Lemma 3.21, we have

$$\begin{aligned}(F(x^k) - F^*)^2 &\leq \left\langle \nabla F(x^k), x^k - x^* \right\rangle^2 \\ &\leq \|\nabla F(x^k)\|^2 \|x^k - x^*\|^2 \leq \beta R(x^0)^2 (F(x^k) - F_{k+1}).\end{aligned}$$

Taking expectation on both sides with respect to ξ_{k-1}, we get

$$(F_k - F^*)^2 \leq \beta R(x^0)^2 (F_k - F_{k+1}).$$

Via the above and by monotonicity of F_k we find that

$$\frac{1}{F_{k+1} - F^*} - \frac{1}{F_k - F^*} = \frac{F_k - F_{k+1}}{(F_k - F^*)(F_{k+1} - F^*)} \geq \frac{1}{\beta R(x^0)^2} \frac{F_k - F^*}{F_{k+1} - F^*} \geq \frac{1}{\beta R(x^0)^2}.$$

Summing terms from 0 to $k-1$ yields

$$\frac{1}{F_k - F^*} - \frac{1}{F(x^0) - F^*} \geq \frac{k}{\beta R(x^0)^2},$$

and, rearranging, we derive

$$F_k - F^* \leq \frac{\beta R(x^0)^2}{k + \beta \frac{R(x^0)^2}{F(x^0)-F^*}}.$$

To eliminate dependence on the starting point, we use Proposition 2.39 *(i)*,

$$F(x^0) - F^* \leq \frac{L}{2}\|x^0 - x^*\|^2 \leq \frac{L}{2}R(x^0)^2,$$

which gives us

$$F_k - F^* \leq \frac{\beta R(x^0)^2}{k + 2\frac{\beta}{L}}.$$

□

3.6.3 The Polyak–Łojasiewicz inequality

The next result states that for L-smooth functions that satisfy the PŁ inequality, we achieve a linear convergence rate. A function is said to satisfy the PŁ inequality with parameter $\mu > 0$ if, for all $x \in \mathbb{R}^n$,

$$\frac{1}{2}\|\nabla F(x)\|^2 \geq \mu \left(F(x) - F^*\right). \tag{3.23}$$

Originally formulated by Polyak in 1963 [178], it was recently shown that this inequality is weaker than other properties commonly used to prove linear convergence [58, 120, 154]. This is useful for extending linear convergence rates to functions that are not strongly convex, including some nonconvex functions.

Proposition 3.26 ([120]). *Let $F : \mathbb{R}^n \to \mathbb{R}$ be μ-convex. Then F satisfies the PŁ inequality* (3.23) *with parameter μ.*

We now proceed to the main result of this subsection.

Theorem 3.27. *Let F be L-smooth and satisfy the PŁ inequality* (3.23) *with parameter μ. Then the three discrete gradient methods and the randomised Itoh–Abe method obtain the linear convergence rate*

$$F_k - F^* \leq \left(1 - \frac{2\mu}{\beta}\right)^k \left(F(x^0) - F^*\right), \tag{3.24}$$

with β given in Table 3.1.

Proof. We combine the PŁ inequality (3.23) with the estimate in Lemma 3.21 to get

$$F(x^k) - F_{k+1} \geq \frac{2\mu}{\beta}(F(x^k) - F^*).$$

By taking expectation of both sides with respect to ξ_{k-1}, we obtain

$$F_{k+1} - F^* \leq \left(1 - \frac{2\mu}{\beta}\right)(F_k - F^*),$$

from which the result follows. □

3.7 Finite path of iterates

In this section, we prove that the *Kurdyka–Łojasiewicz inequality* can be applied to discrete gradient methods to ensure convergence of the iterates $(x^k)_{k\in\mathbb{N}}$ to a unique limit x^*.

The Łojasiewicz inequality was first studied in the context of gradient flows for analytic functions, and is used to prove that if the path of a gradient flow admits an accumulation point, then this point is also the unique limit of the path. Kurdyka later extended this result to functions that are definable in o-minimal structures [126]—for the definition of definable functions, we refer to the same paper. This has since been applied to prove convergence to a unique limit of the iterates of various optimisation algorithms for smooth as well as nonsmooth problems [2, 7, 24, 165, 163].

We recall the result of Kurdyka [126].

Proposition 3.28 (Kurdyka–Łojasiewicz inequality)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable, definable function. Then there is $\varepsilon > 0$, a neighbourhood N_{x^*} of x^*, and a continuously differentiable, strictly increasing function $\psi : [0,\infty) \to (0,\infty)$, such that*

$$\|\nabla F(x)\| \geq \frac{1}{\psi'(F(x) - F(x^*))}, \quad \textit{for all } x \in N_{x^*} \cap \{y \in \mathbb{R}^n \, : \, F(x) - F(x^*) \in (0,\varepsilon)\}. \tag{3.25}$$

Definition 3.29 (Strong Kurdyka–Łojasiewicz inequality)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. We say that the* strong Kurdyka–Łojasiewicz inequality holds for F at $x^* \in \mathbb{R}^n$ *if* (3.25) *holds and ψ is concave.*

We proceed to prove the convergence result, which is a simple application of standard Łojasiewicz arguments.

Theorem 3.30. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a coercive, L-smooth function, and suppose x^* is an accumulation point of the iterates $(x^k)_{k\in\mathbb{N}}$. For each of the four discrete gradient methods, if*

F satisfies the strong Kurdyka–Łojasiewicz inequality at x^*, then

$$\lim_{k\to\infty} x^k = x^*.$$

Proof. In [2, Theorem 3.4], it is proven that if gradient descent-type methods satisfy the growth condition given by

$$C(F(x^k) - F_{k+1}) \geq \|\nabla F(x^k)\| \|x^{k+1} - x^k\|, \tag{3.26}$$

then the existence of an accumulation point x^* at which the strong Kurdyka–Łojasiewicz inequality holds implies that $x^k \to x^*$.

If F is L-smooth, then by (2.9) and (3.18), it follows that (3.26) holds for $C = \sqrt{\beta \tau_{\max}}$. Thus their proof is applicable to the setting of discrete gradient methods, and the result follows. □

3.8 Preconditioned discrete gradient method

We briefly discuss the generalisation of the discrete gradient method (2.8) to a preconditioned version

$$x^{k+1} = x^k - A_k \overline{\nabla} F(x^k, x^{k+1}), \tag{3.27}$$

where $(A_k)_{k\in\mathbb{N}} \subset \mathbb{R}^{n\times n}$ is a sequence of positive-definite matrices. Denoting by $\lambda_{1,k}$ and $\lambda_{n,k}$ the smallest and largest singular values of A_k respectively, we have, for all x,

$$\lambda_{1,k}\|x\| \leq \|A_k x\| \leq \lambda_{n,k}\|x\|.$$

It is straightforward to extend the results in Section 3.3 and Section 3.6 to this setting, under the assumption that there are $\lambda_{\max} \geq \lambda_{\min} > 0$ such that $\lambda_{\min} \leq \lambda_{1,k}, \lambda_{n,k} \leq \lambda_{\max}$ for all $k \in \mathbb{N}$.

There are several possible motivations for this preconditioning. In the context of geometric integration, it is typical to group the gradient flow system (1.12) with the more general dissipative system

$$\dot{x} = -A(x)\nabla F(x),$$

where $A(x) \in \mathbb{R}^{n\times n}$ is positive-definite for all $x \in \mathbb{R}^n$ [183]. This yields numerical schemes of the form (3.27), where we absorb τ_k into A_k. There are optimisation problems in which the time step τ_k should vary for each coordinate. This is, for example, the case when one derives the SOR method from the Itoh–Abe discrete gradient method [153]. More generally, if one has coordinate-wise Lipschitz constants for the gradient of the objective function, it may be beneficial to scale the coordinate-wise time steps accordingly.

3.9 Numerical experiments

In this section, we apply the discrete gradient methods to various test problems. The codes for the figures have been implemented in Python and MATLAB. For solving the discrete gradient equation (2.8) with the Gonzalez and mean value discrete gradients, we use the fixed point method (3.11) detailed in Section 3.4 and tested numerically in Section 3.9.6 under the label '**R**'. For solving (2.8) for the Itoh–Abe method, we use the built-in solver `scipy.optimize.fsolve` in Python.

3.9.1 Setup

We fix the following time steps for the different methods, unless otherwise specified. For the mean value discrete gradient method, we use $\tau_{\mathrm{MV}} = 2/L$, for the Gonzalez discrete gradient method, we use $\tau_{\mathrm{G}} = 2/L$, and for the Itoh–Abe methods, we use the coordinate-dependent time steps $\tau_{\mathrm{IA},i} = \tau_{\mathrm{RIA},i} = 2/L_i$. Note that the time steps for the Itoh–Abe discrete gradient method are not the optimal choice suggested in Table 3.1, but were heuristically optimal for the test problems we considered.

In figure captions and legends, the abbreviations *CIA* and *RIA* refer respectively to the (cyclic) Itoh–Abe discrete gradient method and the randomised Itoh–Abe method drawing uniformly from the standard coordinates. For the sake of comparison, we define one iterate of the randomised Itoh–Abe methods as n scalar updates, so that the computational time is comparable to the standard Itoh–Abe discrete gradient method.

Unless otherwise specified, matrices and vectors for the test problems were created from independent, random, draws from the standard Gaussian distribution in 1D. To provide the matrix with a given condition number, we compute its singular value decomposition and linearly transform its eigenvalues accordingly.

3.9.2 Linear systems

We first solve linear systems of the form

$$\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{2}\|Ax - b\|^2, \tag{3.28}$$

where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.

For linear systems, the Gonzalez and the mean value discrete gradient are both given by

$$\overline{\nabla}F(x,y) = \nabla F\left(\frac{x+y}{2}\right) = A^*\left(A\frac{x+y}{2} - b\right),$$

so we consider these jointly. As discussed previously, the Itoh–Abe methods reduce to SOR methods for solving linear systems and are therefore explicit.

Effect of the condition number

We set $n = 500$ and consider one linear system with a low condition number $\kappa = L/\mu = 10^2$ and one with a high condition number $\kappa = 10^8$. In both cases, we set $x^0 = 0$. See Figure 3.1 for the results for both cases.

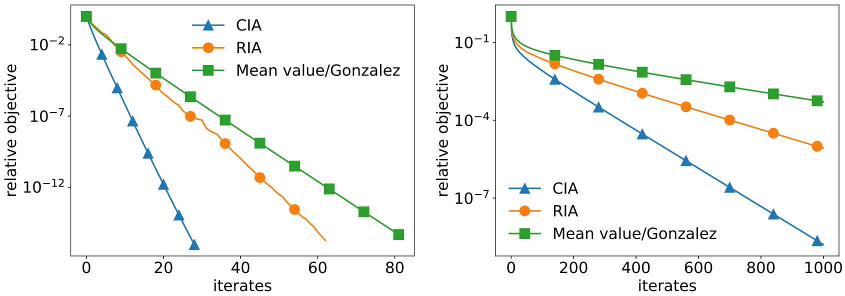


Fig. 3.1 DG methods for linear systems with condition number $\kappa = 10$ (**left**) and $\kappa = 1{,}000$ (**right**). Convergence rate plotted as relative objective $[F(x^k) - F^*]/[F(x^0) - F^*]$. Linear rate is observed for all methods and is sensitive to condition number.

Sharpness of proven convergence rates

We test the sharpness of the convergence rate (3.24)

$$\mathbb{E}_{\xi^{k-1}}[F(x^k)] - F^* \leq \left(1 - \frac{2\mu}{\beta}\right)^k (F(x^0) - F^*),$$

for the randomised Itoh–Abe method. To do so, we run 100 instances of the numerical experiment in the previous subsection and plot the mean convergence rate and 90%-confidence intervals, and compare the results to the proven rate. We do this for two condition numbers, $\kappa = 1.2$ and 10. The results are presented in Figure 3.2. These plots suggest that the proven convergence rate estimate is sharp for the randomised Itoh–Abe method.

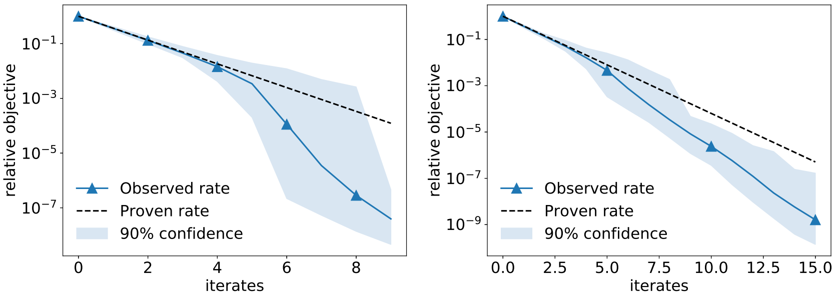


Fig. 3.2 Comparison of observed convergence rate with theoretical convergence rate (3.24), for randomised Itoh–Abe method applied to linear system with condition numbers $\kappa = 1.2$ (**left**) and $\kappa = 10$ (**right**). Average convergence rate and confidence intervals are estimated from 100 runs on the same linear system. The sharpness of the proven convergence rate is observed in both cases.

Linear system with kernel

Next we consider linear systems where the operator A has a nontrivial kernel, meaning that the objective function is not strongly convex, but nevertheless satisfies the PŁ inequality. We let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, where $n = 800$ and $m = 400$, meaning the kernel of A has dimension 400. See Figure 3.3 for the numerical results.

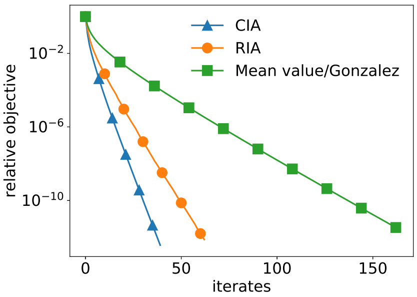


Fig. 3.3 DG methods for linear systems with nontrivial kernel, and convergence rate plotted as relative objective. Due to the kernel, the function is not strongly convex but nevertheless satisfies the PŁ inequality, hence the linear convergence rates.

A note of caution

The performance of coordinate descent methods and their optimal time steps varies significantly with the structure of the optimisation problem [103, 209, 222]. If the linear systems above were constructed with random draws from a distribution whose mean is not zero, then the results would look different. We demonstrate this with a numerical test with results in Figure 3.4.

We compare two time steps for the cyclic Itoh–Abe method, $\tau_i = 2/L_i$ and $\tau_i = 2/(L_i\sqrt{n})$, denoted by the curves labelled “heuristic" and “proven" respectively. While the heuristic time step was superior for most of the test problems considered in this section, it performs significantly worse for this example. Furthermore, in this case the randomised Itoh–Abe method converges faster than the cyclic one.

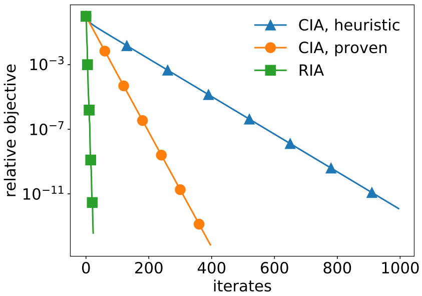


Fig. 3.4 CIA and RIA methods applied to linear system, with matrix entries created from uniform distribution. CIA with the time step $\tau = 1/[\sqrt{n}L]$ **(orange, circle)** performs better than the same method with heuristic time step $\tau = 2/L$ (**blue, triangle**), but worse than RIA. This is the reverse of what is observed if the matrix entries are created from Gaussian distribution.

3.9.3 Regularised logistic regression

We consider a l_2-regularised logistic regression problem, with training data $\left\{x^i, y_i\right\}_{i=1}^m$, where $x^i \in \mathbb{R}^n$ is the data and $y_i \in \{-1,1\}$ is the class label. We wish to solve the optimisation problem

$$\min_{w\in\mathbb{R}^n} F(w) = C\sum_{i=1}^m \log(1+\mathrm{e}^{-y_i\langle w,x^i\rangle}) + \frac{1}{2}\|w\|^2, \tag{3.29}$$

where $C > 0$. We set $n = 100$, $m = 200$, $C = 1$, and the elements of $(y_i)_{i=1}^m$ is drawn from $\{-1,1\}$ with equal probability. The mean value discrete gradient is given by

$$\overline{\nabla}F(w,z) = C\sum_{i=1}^m \frac{\log\left(1+\mathrm{e}^{-y_i\langle x^i,w\rangle}\right) - \log\left(1+\mathrm{e}^{-y_i\langle x^i,z\rangle}\right)}{\langle x^i, w-z\rangle} x^i + \frac{w+z}{2},$$

and the Gonzalez discrete gradient is given by

$$\overline{\nabla} F(w,z) = C \sum_{i=1}^{m} \left(\frac{-y_i \mathrm{e}^{-y_i \left\langle \frac{w+z}{2}, x^i \right\rangle}}{1+\mathrm{e}^{-y_i \left\langle \frac{w+z}{2}, x^i \right\rangle}} \left(x^i - \frac{\langle x^i, w-z\rangle}{\|w-z\|^2}(w-z) \right) \right.$$
$$\left. + \frac{\log(1+\mathrm{e}^{-y_i\langle w,x^i\rangle}) - \log(1+\mathrm{e}^{-y_i\langle z,x^i\rangle})}{\|w-z\|^2}(w-z) \right) + \frac{w+z}{2}.$$

See Figure 3.5 for the numerical results.

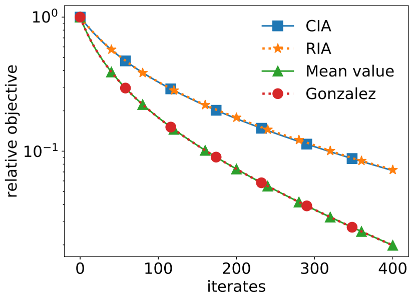


Fig. 3.5 DG methods for l_2-regularised logistic regression. Convergence rate plotted as relative objective. The rates of randomised and cyclic Itoh–Abe methods almost coincide, and so do the mean value and Gonzalez discrete gradient methods.

3.9.4 Nonconvex function

We solve the nonconvex problem

$$\min_{x\in\mathbb{R}^n} F(x) = \|Ax\|^2 + 3\sin^2(\langle c,x\rangle), \tag{3.30}$$

where $A \in \mathbb{R}^{n\times n}$ is a square, nonsingular matrix, and $c \in \mathbb{R}^n$ satisfies $Ac = c$ and $\|c\| = 1$. This is a higher-dimensional extension of the scalar function $x^2 + 3\sin^2(x)$ considered by Karimi et al. in [120]. This scalar function satisfies the PŁ inequality (3.23) for $\mu = 1/32$, and it follows that F satisfies it for $\mu = 1/(32\kappa)$, where κ is the condition number of A^*A. Furthermore, the nonconvexity of F can be observed by considering the restriction of F to $x = \lambda c$ for $\lambda \in \mathbb{R}$, which has the form of the original scalar function. The function has the unique minimiser $x^* = 0$ with $F^* = 0$.

We set $n = 50$ and choose x^0 constructed by random, independent draws from a Gaussian distribution. See Figure 3.6 for the numerical results.

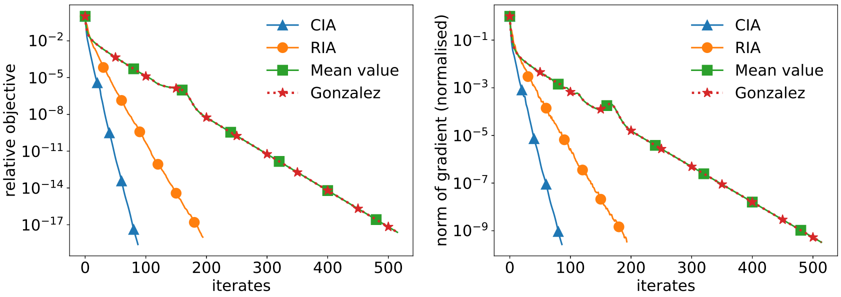


Fig. 3.6 DG methods applied to nonconvex problem that satisfies the PŁ inequality. **Left**: Plots of relative objective. **Right**: Plots of norm of gradient (normalised) $\|\nabla F(x^k)\| \, / \, \|\nabla F(x^0)\|$.

3.9.5 Comparison of Itoh–Abe and explicit coordinate descent for stiff problems

As discussed in Chapter 1, variational optimisation problems for image analysis and signal processing often feature nonsmooth regularisation terms that promote sparsity, e.g. in the gradient domain or a wavelet basis. To overcome the nonsmoothness, these terms may be replaced with smooth approximations. This, however, leads to optimisation problems that suffer from stiffness, i.e. local, rapid variations in the gradient, requiring the use of severely small time steps for explicit numerical methods. In such cases, the cost of solving an implicit equation such as (2.8) may be preferrable to explicit methods.

We investigate this scenario, by comparing the Itoh–Abe discrete gradient method to explicit coordinate descent, for solving (smoothened) total variation denoising problems. We consider a ground truth image $x^{\text{true}} \in \mathbb{R}^n$ and a noisy image $x^\delta = x^{\text{true}} + \delta$, where δ is random Gaussian noise. The total variation regulariser is defined as $\text{TV}(x) := \sum_{i=1}^n |[\nabla x]_i|$, with $\nabla : \mathbb{R}^n \to \mathbb{R}^{2\times n}$ a discretised spatial gradient as defined in [47], and $|\cdot| : \mathbb{R} \to \mathbb{R}$ the absolute value function. As the nonsmoothness is induced by the absolute value function, we approximate the regulariser by $TV_\varepsilon(x) := \sum_{i=1}^n |[\nabla x]_i|_\varepsilon$, where

$$|x|_\varepsilon := \sqrt{x^2 + \varepsilon}.$$

The optimisation problem is thus given by

$$\operatorname*{arg\,min}_{x\in\mathbb{R}^n} \frac{1}{2}\|x - x^\delta\|^2 + \lambda \, \text{TV}_\varepsilon(x). \tag{3.31}$$

Unless otherwise specified, the time step for explicit coordinate descent (CD) is $\tau_{\text{CD}} = 1/(2\lambda\sqrt{\varepsilon}+1)$ and for the Itoh–Abe discrete gradient method (DG) is $\tau_{\text{DG}} = 1/10$.

In Figure 3.7, we compare the DG method for a range of time steps to CD. This demonstrates that the superior convergence rate of the DG method is stable with respect to a wide range of time steps. In Figure 3.8, we compare the DG method to CD for different values of ε, demonstrating that the benefits of using the DG method increases as ε gets smaller. In Figure 3.9, we compare different time steps for CD to the DG method, showing that for large time steps, the scheme is unstable and fails to decrease while for small time steps, the iterates decrease too slowly. In Figure 3.10, we employ a simple backtracking line search (LS) method based on the Armijo-Goldstein condition, and compare this to the DG method.

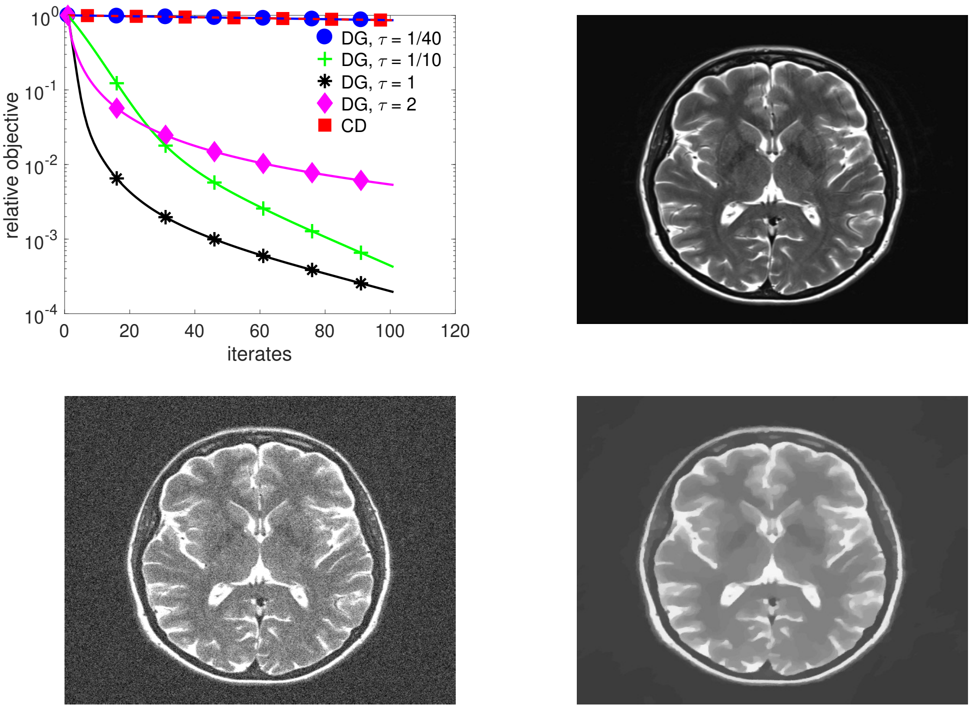


Fig. 3.7 Top left: Comparison of explicit coordinate descent with $\tau_{\text{CD}} = 1/(2\lambda\sqrt{\varepsilon}+1)$ vs the Itoh–Abe discrete gradient methods with time steps 0.025, 0.1, 1, and 2, and with $\varepsilon = 10^{-8}$. Top right: Ground truth image. Bottom left: Noisy image. Bottom right: Total variation denoising with $\varepsilon = 10^{-8}$.

3.9.6 Comparison of methods for solving the discrete gradient equation

We test the numerical performance of four methods for solving the discrete gradient equation (2.8), building on the fixed point theory in Section 3.4.

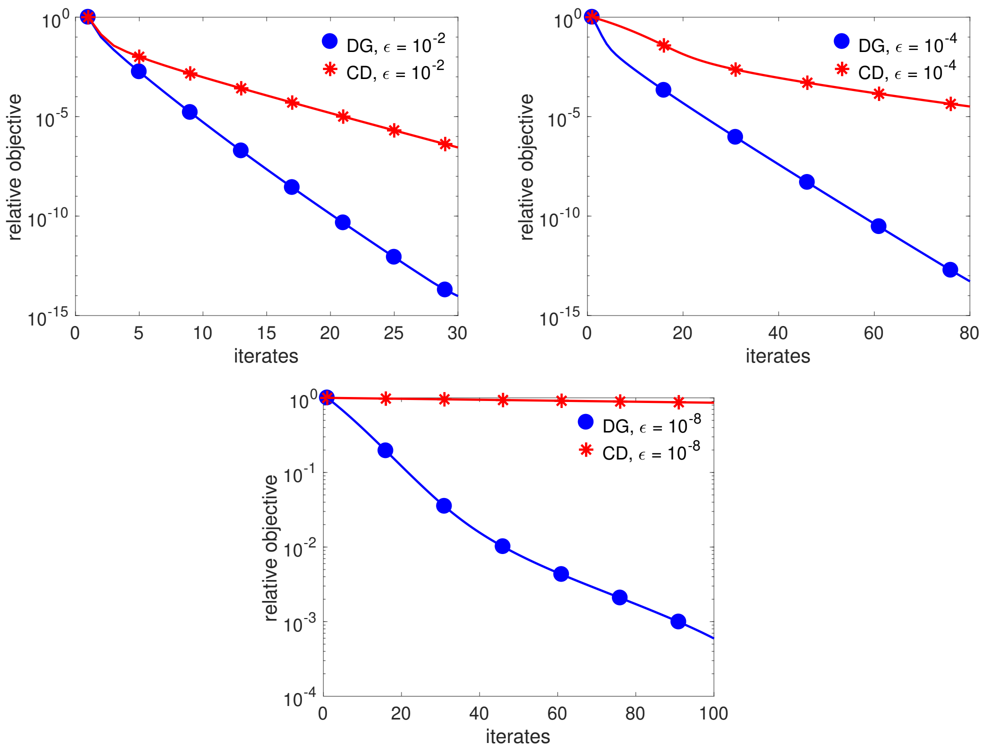


Fig. 3.8 Comparison of CD to DG for three values of ε, 10^{-2}, 10^{-4}, and 10^{-8}. The time steps are set to $\tau_{\mathrm{DG}} = \sqrt{\tau_{\mathrm{CD}}}$ where the latter time step is set to $1/L$.

The first method, denoted **F**, is the fixed point updates (3.10) proposed in [164] ($\theta = 1$). The second method, denoted **R**, is the relaxed fixed point method (3.11), where θ is optimised according to (3.12) if F is convex, and is otherwise set to $1/2$. The third method, denoted **F+R**, is also the updates (3.11) with $\theta = 1$ by default, but whenever the discrepancy $\|T(y^{k+1}) - y^{k+1}\|$ is greater than $\|T(y^k) - y^k\|$, then the update is repeated with θ set to half its previous value. This third option might be desirable in cases where $\theta = 1$ is expected to give faster convergence but also be unstable. The fourth method is the built-in solver `scipy.optimize.fsolve` in Python.

To test these methods, we performed 50 iterations of the discrete gradient method for different test problems, where at each iterate the discrete gradient solver would run until

$$\|r^k\|_\infty < \varepsilon, \quad \text{where } r_i^k := \frac{y_i^k - y_i^{k-1}}{y_i^{k-1}} \text{ if } y_i^{k-1} \neq 0, \text{ and } r_i^k := y_i^k \text{ otherwise,}$$

for a specified tolerance $\varepsilon > 0$, or until k reaches a given maximum $K_{\max}$. We then compare the average CPU time (s) for each of these methods. If a method fails to converge to a fixed

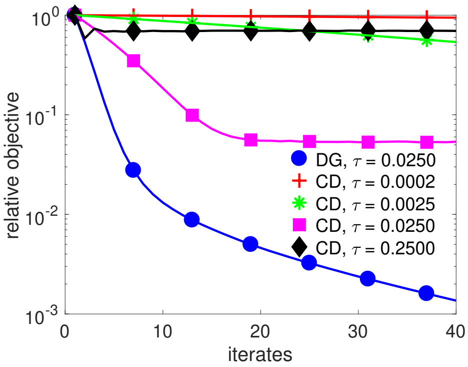


Fig. 3.9 Comparison of different time steps for CD vs fixed time step for DG. For smaller time steps, the CD iterates decrease too slowly, and for larger steps, they become unstable and fail to decrease.

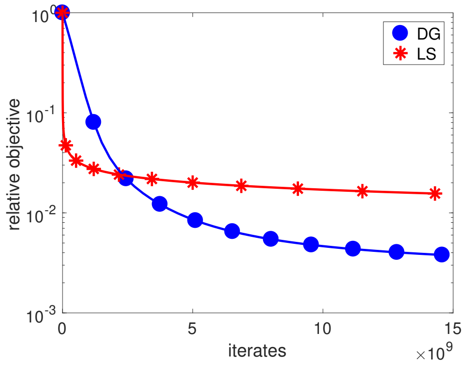


Fig. 3.10 Comparison of DG to simple backtracking line search (LS) in terms of coordinate evaluations.

point for a significant number of the iterations ($> 10\%$), we consider the method inapplicable for that test problem.

We test the methods for the mean value discrete gradient applied to three of the previous test problems, for $\varepsilon = 10^{-6}$ and 10^{-12}. We have not included results for the Gonzalez discrete gradient and other tolerances, as the results were largely the same.

The results are given in Table 3.2. We see that **R** is superior in stability, being the only method that locates the minimiser in every case. In all cases, **R** or **F+R** were the most efficient or close to the most efficient method. However, the relative performance of the different methods varies notably for the different test problems. This suggests that optimising for θ would require it to be tuned according to the optimisation problem, e.g. by an initial line search procedure.

Table 3.2 Average CPU time (s) over 50 iterations of (2.8) with the mean value discrete gradient. Tolerance $\varepsilon = 10^{-6}$.

Test problem	F	R	F + R	fsolve	ε
Linear system (3.28)	N/A (0.003)	0.006	**0.002**	0.190	10^{-6}
Logistic regression (3.29)	**0.001**	0.016	0.001	N/A (0.054)	
Nonconvex problem (3.30)	N/A (0.019)	**0.003**	N/A (0.020)	N/A (0.427)	
Linear system (3.28)	N/A (0.011)	0.012	**0.005**	0.206	10^{-12}
Logistic regression (3.29)	0.055	0.037	**0.019**	N/A (0.076)	
Nonconvex problem (3.30)	N/A (0.033)	**0.005**	N/A (0.031)	0.513	

3.10 Conclusion and outlook

In this chapter, we have studied the discrete gradient method for optimisation, and provided several fundamental results on well-posedness, convergence rates and optimal time steps. We have focused on four methods, using the Gonzalez discrete gradient, the mean value discrete gradient, the Itoh–Abe discrete gradient, and a randomised version of the Itoh–Abe method. Several of the proven convergence rates match the optimal rates of classical methods such as gradient descent and stochastic coordinate descent. For the Itoh–Abe discrete gradient method, the proven rates are better than previously established rates for comparable methods, i.e. cyclic coordinate descent methods [221].

There are open problems to be addressed in future work. First, similar to acceleration for gradient descent and coordinate descent [15, 157, 159, 221], we will study acceleration of the discrete gradient method to improve the convergence rate from $O(1/k)$ to $O(1/k^2)$.

Chapter 4

Discrete gradient methods for nonsmooth, nonconvex optimisation

4.1 Introduction

This chapter is based on the preprint [184], and is joint work with Matthias J. Ehrhardt, G. R. W. Quispel, and Carola-Bibiane Schönlieb.

In the previous chapter, we studied and provided analyis for discrete gradient methods in the continuously differentiable setting. In this chapter, we switch the focus to nonsmooth, nonconvex optimisation problems.

Thus we consider the unconstrained problem

$$\min_{x \in \mathbb{R}^n} F(x), \tag{4.1}$$

where the objective function F is locally Lipschitz continuous, bounded below and coercive. The function may be nonconvex and nonsmooth, and we assume no knowledge besides point evaluations $x \mapsto F(x)$. To solve (4.1), we consider generalised Itoh–Abe type methods, namely the randomised Itoh–Abe methods studied in Chapter 3, as well as a deterministic variant. In this chapter, we therefore seek to extend discrete gradient methods from the differentiable setting to the nonsmooth setting.

Itoh–Abe methods

We recall the Itoh–Abe scalar update (3.4), defined via

$$x^{k+1} \mapsto x^k - \tau_k \alpha_k d^{k+1}, \quad \text{where } \alpha_k \neq 0 \text{ solves} \quad \alpha_k = -\frac{F(x^k - \tau_k \alpha_k d^{k+1}) - F(x^k)}{\tau_k \alpha_k}.$$

We thus refer to α_k as the implicit solution to this scalar equation, and consider the following algorithm.

Algorithm 1 Generalised Itoh–Abe method

Input: starting point x^0, directions $(d^k)_{k\in\mathbb{N}}$, time steps $(\tau_k)_{k\in\mathbb{N}}$.

```
for $k = 0, 1, 2, \ldots$ do
    Update $x^{k+1} = x^k - \tau_k \alpha_k d^{k+1}$ via (3.4)
end for
```

4.1.1 Bilevel optimisation and blackbox problems

An important motivation for the methods studied in this chapter is nonsmooth bilevel problems, which we introduced in Chapter 1. We briefly recall these problems in the more general setting of simulation-based optimisation. We suppose a simulation model depends on some tunable parameters $\vartheta \in \mathbb{R}^n$, such that for a given parameter choice ϑ, the model returns an output x_ϑ. Furthermore, there is a cost function Φ, which assigns to output x_ϑ a numerical score $\Phi(x_\vartheta) \in \mathbb{R}$, which we want to minimise with respect to ϑ. The associated parameter optimisation problem becomes

$$\vartheta^* \in \arg\min_{\vartheta\in\mathbb{R}^n} \Phi(x_\vartheta).$$

Another example of parameter optimisation problems is supervised machine learning.

In this chapter, we consider bilevel problems for variational regularisation models, i.e. (1.11). Namely, we consider a variational regularisation problem for image denoising,

$$x_\vartheta \in \arg\min_x \frac{1}{2}\|x - f^\delta\|^2 + R_\vartheta(x),$$

where f^δ is a noisy image and ϑ is the regularisation parameter. For training data with desired reconstruction $x^\dagger$, we consider a scoring function Φ that estimates the discrepancy between $x^\dagger$ and the reconstruction x_ϑ. In Section 4.5.2, we apply generalised Itoh-Abe methods to solve these problems.

As discussed in Section 1.1.2, bilevel problems, and parameter optimisation problems in general, pose several challenges. They are often nonconvex and nonsmooth, due to the nonsmoothness and nonlinearity of $\vartheta \mapsto x_\vartheta$. Furthermore, the model simulation $\vartheta \mapsto x_\vartheta$ is an algorithmic process for which gradients or subgradients are challenging to compute[1].

[1] Note that we address this issue for bilevel problems in depth in Chapter 7.

Such problems can then be modelled as *blackbox optimisation problems*, for which one only has access to point evaluations of the function. It is therefore of great interest to develop efficient and robust derivative-free methods for such optimisation problems.

There is a rich literature on bilevel optimisation for variational regularisation problems in image analysis, c.f. e.g. [40, 66, 125, 166]. In Chapter 7 we provide a wider literature review for this topic.

Furthermore, model parameter optimisation problems appear in many other applications. These include optimising for the management of water resources [91], approximation of a transmembrane protein structure in computational biology [98], image registration in medical imaging [168], the building of wind farms [77], and solar energy utilisation in architectural design [119], to name a few.

4.1.2 Related literature on nonsmooth, nonconvex optimisation

Although nonsmooth, nonconvex problems are known for their difficulty compared to convex problems, a rich optimisation theory has grown since the 1970s. As the focus of this chapter is derivative-free optimisation, we will compare the methods' convergence properties and performance to other derivative-free solvers. Audet and Hare [10] recently provided a reference work for this field.

While there is a myriad of derivative-free solvers, few provide convergence guarantees for nonsmooth, nonconvex functions. Audet and Dennis Jr [9] introduced the *mesh adaptive direct search* (MADS) method for constrained optimisation, with provable convergence guarantees to stationary points for nonsmooth, nonconvex functions in the Clarke subdifferential framework. Direct search methods evaluate the function at a finite polling set, compare the evaluations, and update the polling set accordingly. Such methods only consider the ordering of the evaluations, rather than the numerical differences. A significant portion of derivative-free methods are direct search methods, and the most well-known of these is the Nelder–Mead method (also known as the downhill simplex method) [155].

Alternatively, derivative-free model-based methods that build a local quadratic model based on evaluations are well-documented [42, 180, 181]. While such methods tend to work well in practice, they are normally designed only for smooth functions, so their performance on nonsmooth functions is not guaranteed.

Fasano et al. [87] formulated a derivative-free line search method termed DFN and analyse its convergence properties for nonsmooth functions for the Clarke subdifferential, in the constrained setting. Building on the DFN algorithm, Liuzzi and Truemper [139] formulated a derivative-free method that is a hybrid between DFN and MADS. The Itoh–Abe methods share many similarities with DFN, such as performing line searches along dense directions, and they employ a similar convergence analysis. However, the line search methods differ, and the Itoh–Abe methods are in particular motivated by structure-preservation of gradient flow-type dissipativity. Furthermore, our convergence analysis is more comprehensive, considering both stochastic and deterministic methods, and obtaining convergence guarantees using the cyclical density property.

Furthermore, we note the resemblance of randomised Itoh–Abe methods (3.4), when $(d^k)_{k\in\mathbb{N}}$ is randomly, independently drawn from S^{n-1}, to the random search method proposed by Polyak in [179] and studied for nonsmooth, convex functions by Nesterov in [160], given by

$$x^{k+1} = x^k - \tau_k \frac{F(x^k + \alpha_k d^{k+1}) - F(x^k)}{\alpha_k} d^{k+1},$$

where the sequence $(d^k)_{k\in\mathbb{N}}$ is randomly, independently drawn from S^{n-1}. The implicit equation (3.4) can be treated as a line search rule for the above method, with constraints imposed by $\tau_{\min}$, $\tau_{\max}$.

While our focus is on derivative-free methods, we also mention some popular methods for nonsmooth, nonconvex optimisation that use gradient or subgradient information. Central in nonsmooth optimisation are *bundle methods*, where a subgradient [54] is required at each iterate to construct a linear approximation to the objective function—see [121] for an introduction. A close alternative to bundle methods are *gradient sampling methods* (see [38] for a recent review by Burke et al.), where the descent direction is determined by sampling gradients in a neighbourhood of the current iterate. Curtis and Que [60] formulated a hybrid method between the gradient sampling scheme of [59] and the well-known quasi-Newton method BFGS adapted for nonsmooth problems [133]. These methods have convergence guarantees in the Clarke subdifferential framework, under the assumption that the objective function is differentiable in an open, dense set. Last, we mention a derivative-free scheme based on gradient sampling methods, proposed by Kiwiel [123], where gradients are replaced by Gupal's estimates of gradients of the Steklov averages of the objective function. This method has convergence guarantees in the Clarke subdifferential framework, but has a high computational cost in terms of function evaluations per iterate.

4.1.3 Contributions

In this chapter, we formulate generalised Itoh–Abe methods for solving nonsmooth functions. We prove that the methods always admit a solution, and that the iterates converge to a set of Clarke stationary points, for any locally Lipschitz continuous function, and both for deterministic and randomly chosen search directions. Consequently, the scope of discrete gradient methods for optimisation is significantly broadened, and we conclude that the dissipativity properties of gradient flows can meaningfully be preserved even beyond differentiability. Ultimately, this provides a new, robust, and versatile optimisation scheme for nonsmooth, nonconvex functions.

The theoretical convergence analysis for the Itoh–Abe methods is thorough and foundational, and we provide examples that demonstrate that the conditions of the convergence theorem are not just sufficient, but necessary. Furthermore, the statements and proofs are sufficiently general so that they can be adapted to other schemes, such as the aforementioned DFO method, thus enhancing the theory of these methods as well.

We show that the method works well in practice, by solving bilevel optimisation problems for variational regularisation problems, as well as solving benchmark problems such as Rosenbrock functions.

The rest of the chapter is structured as follows. In Section 4.2, the main theoretical results of the chapter are presented, namely existence and optimality results in the stochastic and deterministic setting. In Section 4.3, we briefly discuss the Itoh–Abe discrete gradient for general coordinate systems. In Section 4.4 and Section 4.5, the numerical implementation is described and results from test problems are presented.

4.2 The discrete gradient method for nonsmooth optimisation

In this section, we present the main theoretical results for the generalised Itoh–Abe methods. In particular, we prove that the update (3.4),

$$x^{k+1} \mapsto x^k - \tau_k \alpha_k d^{k+1}, \quad \text{where } \alpha_k \neq 0 \text{ solves} \quad \alpha_k = -\frac{F(x^k - \tau_k \alpha_k d^{k+1}) - F(x^k)}{\tau_k \alpha_k},$$

admits a solution for all $\tau_k > 0$. We also prove under minimal assumptions on F and $(d^k)_{k\in\mathbb{N}}$ that the iterates converge to a connected set of Clarke stationary points, both in a stochastic and deterministic setting.

4.2.1 Existence result

Lemma 4.1. *Suppose F is a continuous function bounded below, and that $x \in \mathbb{R}^n$, $d \in S^{n-1}$ and $\tau > 0$. Then at least one of the following statements hold.*

(i) There is $\alpha \neq 0$ that solves (3.4), i.e. that satisfies $\frac{F(x-\tau\alpha d)-F(x)}{\tau\alpha} = -\alpha$.

(ii) F is Clarke directionally stationary at x along d.

Proof. Suppose the second statement does not hold. Then there is $\varepsilon > 0$ such that

$$\min\left\{F^o(x;-d), F^o(x;d)\right\} < -\varepsilon,$$

so assume without loss of generality that $F^o(x;-d) < -\varepsilon$. By definition of F^o, there is $\delta > 0$ such that for all $\alpha \in (0,\delta)$,

$$\frac{F(x-\tau\alpha d)-F(x)}{\tau\alpha} \leq -\varepsilon/2.$$

Taking $\alpha \to 0$, we get that

$$\lim_{\alpha\to 0^+}\frac{F(x-\tau\alpha d)-F(x)}{\tau\alpha^2} \leq \lim_{\alpha\to 0^+} -\frac{\varepsilon}{2\alpha} = -\infty,$$

so there is a $\alpha_1 \in (0,\delta)$ such that

$$\frac{F(x-\tau\alpha_1 d)-F(x)}{\tau\alpha_1^2} < -1.$$

On the other hand, as F is bounded below, we have

$$\liminf_{\alpha\to\infty}\frac{F(x-\tau\alpha_2 d)-F(x)}{\tau\alpha_2^2} \geq \lim_{\alpha_2\to\infty}\frac{\min\left\{0, F(x-\tau\alpha_2 d)\right\}-F(x)}{\tau\alpha_2^2} = 0.$$

Thus there is α_2 such that

$$\frac{F(x-\tau\alpha_2 d)-F(x)}{\tau\alpha_2^2} > -1.$$

Since the mapping $\alpha \mapsto \frac{F(x-\tau\alpha d)-F(x)}{\tau\alpha^2}$ is continuous for $\alpha \in (0,\infty)$, we conclude by the intermediate value theorem [197, Theorem 4.23] that there is $\alpha \in (\alpha_1,\alpha_2)$ that solves the discrete gradient equation

$$\frac{F(x-\tau\alpha d)-F(x)}{\tau\alpha^2} = -1.$$

□

Remark 4.2. *Note that by the above proof it is straightforward to identify an interval in which a solution to* (3.4) *exists, allowing for the use of standard root solver algorithms.*

The following lemma, which is an adaptation of [100, Theorem 1] for the nonsmooth setting, summarises some useful properties of the methods.

Lemma 4.3. *Suppose that F is continuous, bounded from below and coercive, and let $(x^k)_{k\in\mathbb{N}}$ be the iterates produced by* (3.4)*. Then, the following properties hold.*

(i) $F(x^{k+1}) \leq F(x^k)$.

(ii) $\lim_{k\to\infty} \frac{F(x^{k+1})-F(x^k)}{\|x^{k+1}-x^k\|} = 0$.

(iii) $\lim_{k\to\infty} \|x^k - x^{k+1}\| = 0$.

(iv) $(x^k)_{k\in\mathbb{N}}$ *has an accumulation point* x^*.

Proof. Property *(i)* follows from the equation $F(x^{k+1}) - F(x^k) = -\tau_k \alpha_k^2$.

Next we show properties *(ii)* and *(iii)*. Since F is bounded below and $(F(x^k))_{k\in\mathbb{N}}$ is decreasing, $F(x^k) \to F^*$ for some limit F^*. Therefore, by (3.6)

$$F(x^0) - F^* = \sum_{k=0}^{\infty} F(x^k) - F(x^{k+1}) = \sum_{k=0}^{\infty} \tau_k \left(\frac{F(x^k) - F(x^{k+1})}{\|x^{k+1} - x^k\|} \right)^2$$
$$\geq \tau_{\min} \sum_{k=0}^{\infty} \left(\frac{F(x^k) - F(x^{k+1})}{\|x^{k+1} - x^k\|} \right)^2 .$$

Similarly, by (3.6)

$$F(x^0) - F^* = \sum_{k=0}^{\infty} F(x^k) - F(x^{k+1}) = \sum_{k=0}^{\infty} \frac{1}{\tau_k} \|x^k - x^{k+1}\|^2 \geq \frac{1}{\tau_{\max}} \sum_{k=0}^{\infty} \|x^k - x^{k+1}\|^2.$$

We conclude

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

which proves properties *(ii)* and *(iii)*.

Last, we prove that property *(iv)* holds. Since $(F(x^k))_{k\in\mathbb{N}}$ is a decreasing sequence, the iterates $(x^k)_{k\in\mathbb{N}}$ belong to the set $\left\{x \in \mathbb{R}^n \,:\, F(x) \leq F(x^0)\right\}$. Therefore, by coercivity of F, the iterates $(x^k)_{k\in\mathbb{N}}$ are bounded, and admit an accumulation point. □

We denote by S the limit set of $(x^k)_{k\in\mathbb{N}}$, which is the set of accumulation points,

$$S = \left\{ x^* \in \mathbb{R}^n \,:\, \exists (x^{k_j})_{j\in\mathbb{N}} \text{ s.t. } x^{k_j} \to x^* \right\}. \tag{4.2}$$

By the above lemma, S is nonempty. We now prove further properties of the limit set.

Lemma 4.4. *The limit set S is compact, connected and has empty interior. Furthermore, F is constant on S.*

Proof. Boundedness of S follows from coercivity of F combined with the fact that S belongs to $\{x \in \mathbb{R}^n \,:\, F(x) \leq F(x^0)\}$. Since any accumulation point of S is also an accumulation point of $(x^k)_{k\in\mathbb{N}}$, S is closed. Hence S is compact.

We prove connectedness by contradiction. Suppose there are two disjoint and nonempty open sets A and B such that $S \subset A \cup B$. The sequence $(x^k)_{k\in\mathbb{N}}$ jumps between A and B infinitely many times and $\|x^{k+1} - x^k\| \to 0$, which implies that there is a subsequence of $(x^{k_j})_{j\in\mathbb{N}}$ in $\mathbb{R}^n \setminus (A \cup B)$. However, $(x^{k_j})_{j\in\mathbb{N}}$ is a bounded sequence and has an accumulation point, which must belong in $\mathbb{R}^n \setminus (A \cup B)$. This contradicts the assumption that all accumulation points of $(x^k)_{k\in\mathbb{N}}$ are in $A \cup B$.

We show that S has empty interior by contradiction. Suppose S contains an open ball $B_\varepsilon(x)$ in $\mathbb{R}^n$. Then as $\|x^{k+1} - x^k\| \to 0$, there is a $j \in \mathbb{N}$ such that $x^j \in B_\varepsilon(x) \subset S$. However, as F takes the same value on all of S, we deduce that $F(x^j) = \lim_{k\to\infty} F(x^k)$. Since $(F(x^k))_{k\in\mathbb{N}}$ is a decreasing sequence, $F(x^k) = F(x^j)$ for all $k > j$. It follows from (3.6) that $x^k = x^j$ for all $k > j$. Therefore, $S = \{x^j\}$, which contradicts the assumption that S has nonempty interior.

Last, since $(F(x^k))_{k\in\mathbb{N}}$ is a decreasing sequence and $F(x^*) = \lim_{k\to\infty} F(x^k)$ for all $x^* \in S$, it follows that F is constant on S. □

Discrete gradients versus subgradients

One could hope that the consistency property (2.6) extends to nonsmooth functions, i.e. that the Itoh–Abe discrete gradient converges to a subgradient, and that one could thereby prove that limit points are Clarke stationary. We provide a counterexample to show that this is not the case. That is, for nondifferentiable F, discrete gradients do not necessarily approximate a subgradient or even an ε-*approximate subgradient*.[2]

Example 4.5. *Let* $F(x_1, x_2) := \sqrt{x_1^2 + x_2^2}$*, and set* $x^k = [\frac{1}{k}, 0]^T$ *and* $y^k = [0, \frac{1}{k}]^T$*. We have*

$$\overline{\nabla} F(x^k, y^k) = [1,1]^T, \quad \lim_{k\to\infty} x^k = \lim_{k\to\infty} y^k = [0,0]^T.$$

However, $[1,1]^T$ *is not in* $\partial F(0,0) = B_1(0,0)$*. In fact, for all* $\varepsilon > 0$*, we have* $[1,1]^T \notin \partial_\varepsilon F(0,0)$*.*

4.2.2 Optimality result

We now proceed to the main result of this chapter, namely that all points in the limit set S are Clarke stationary. We consider the stochastic case and the deterministic case separately.

In the stochastic case, we assume that the directions $(d^k)_{k\in\mathbb{N}}$ are randomly, independently drawn, and that the support of the probability density of Ξ is dense in S^{n-1}. It is straightfor-

[2] For convex functions, $p \in \mathbb{R}^n$ is an ε-approximate subgradient if for all $y \in \mathbb{R}^n$ one has $F(y) \geq F(x) + \langle p, y - x \rangle - \varepsilon$ [110].

ward to extend the proof to the case where $(d^{nk+1},\dots,d^{n(k+1)})$ are drawn as an orthonormal system under the assumptions that the directions $(d^{nk+1})_{k\in\mathbb{N}}$ are independently drawn from S^{n-1} and that the support of the density of the corresponding marginal distribution is dense in S^{n-1}.

We define X to be the set of nonstationary points,

$$X = \{x \in \mathbb{R}^n \,:\, 0 \notin \partial F(x)\}. \tag{4.3}$$

Theorem 4.6. *Let $(x^k)_{k\in\mathbb{N}}$ solve* (3.4) *where $(d^k)_{k\in\mathbb{N}}$ are independently drawn from the random distribution Ξ, and suppose that the support of the density of Ξ is dense in S^{n-1}. Then $\mathbb{P}(S\cap X \neq \varnothing) = 0$, i.e. the limit set S is almost surely in the set of stationary points.*

Proof. We will construct a countable collection of open sets $(B_j)_{j\in\mathbb{N}}$, such that $X \subset \bigcup_{j\in\mathbb{N}} B_j$ and so that for all $j\in\mathbb{N}$ we have $\mathbb{P}(S\cap B_j \neq \varnothing) = 0$. Then the result follows from countable additivity of probability measures.

First, we show that for every $x\in X$, there is $d\in S^{n-1}$, $\varepsilon>0$, and $\delta>0$ such that

$$\frac{F(y-\lambda e)-F(y)}{\lambda} \le -\varepsilon, \quad \forall y\in B_\delta(x),\ e\in B_\delta(d)\cap S^{n-1},\ \lambda\in(0,\delta). \tag{4.4}$$

To show this, note that if $x\in X$, then by definition there is $d\in S^{n-1}$ and $\varepsilon>0$ such that

$$F^o(x;-d) = \limsup_{\substack{y\to x\\ \lambda\downarrow 0}} \frac{F(y-\lambda d)-F(y)}{\lambda} \le -\varepsilon.$$

Therefore, there is $\eta>0$ such that for all $\lambda\in(0,\eta)$ and all $y\in B_\eta(x)$, we have

$$\frac{F(y-\lambda d)-F(y)}{\lambda} \le -\varepsilon/2.$$

As F is Lipschitz continuous around $B_\eta(x)$, it is clear that the mapping

$$e \mapsto \frac{F(y-\lambda e)-F(y)}{\lambda},$$

is also locally Lipschitz continuous (of the same rank). It follows that there exists $\delta\in(0,\eta)$ such that for all $y\in B_\delta(x)$, all $e\in B_\delta(d)\cap S^{n-1}$, and all $\lambda\in(0,\delta)$, we have

$$\frac{F(y-\lambda e)-F(y)}{\lambda} \le -\varepsilon/3.$$

This concludes the first part.

Next, for $m \in \mathbb{N}$, we define the set

$$X_m = \left\{ x \in X \ : \ \text{(4.4) holds for some } d \in S^{n-1}, \varepsilon > 0 \text{ and all } \delta < 1/m \right\}.$$

Clearly

$$X = \bigcup_{m \in \mathbb{N}} X_m.$$

Let $(y^i)_{i \in \mathbb{N}}$ be a dense sequence in X_m, which exists because $\mathbb{Q}^n$ is both countable and dense in $\mathbb{R}^n$. We define $Y_i^{(m)} = B_\delta(y^i)$, where $\delta = \frac{1}{m+1}$. Therefore,

$$X_m \subset \bigcup_{i \in \mathbb{N}} Y_i^{(m)} \quad \Longrightarrow \quad X \subset \bigcup_{m \in \mathbb{N}} \bigcup_{i \in \mathbb{N}} Y_i^{(m)}.$$

Since a countable union of countable sets is countable, we conclude with the following statement. For each $i \in \mathbb{N}$ there is $y^i \in \mathbb{R}^n$, $\varepsilon_i, \delta_i > 0$, and $\widetilde{d^i} \in S^{n-1}$, such that for all $z \in B_{\delta_i}(y^i)$, all $\widetilde{d} \in B_{\delta_i}(\widetilde{d^i}) \cap S^{n-1}$, and all $\lambda \in (0, \delta_i)$, we have

$$\frac{F(z - \lambda \widetilde{d}) - F(z)}{\lambda} \leq -\varepsilon_i,$$

and such that

$$X \subset \bigcup_{i \in \mathbb{N}} B_{\delta_i}(y^i). \tag{4.5}$$

Finally, we show that for each $i \in \mathbb{N}$, almost surely, $S \cap B_{\delta_i}(y^i) = \varnothing$. For a given i, write $B_i := B_{\delta_i}(y^i)$, and define $m := \min_{x \in B_i} F(x)$, $M := \max_{x \in B_i} F(x)$. We argue accordingly: The existence of an accumulation point of $(x^k)_{k \in \mathbb{N}}$ in B_i would imply that there is a subsequence $(x^{k_j})_{j \in \mathbb{N}} \subset B_i$. Suppose $x^{k_j} \in B_i$ and $d^{k_j+1} \in B_{\delta_i}(\widetilde{d^i})$, so that $x^{k_j+1} = x^{k_j} - \lambda d^{k_j+1}$ for some $\lambda > 0$. If $\lambda < \delta_i$, then

$$F(x^{k_j} - \lambda d^{k_j+1}) - F(x^{k_j}) \leq -\varepsilon_i \lambda = -\varepsilon_i \| x^{k_j+1} - x^{k_j} \|.$$

However,

$$F(x^{k_j+1}) - F(x^{k_j}) = -\frac{1}{\tau_{k_j}} \| x^{k_j+1} - x^{k_j} \|^2,$$

so, combining these equations, we get

$$\varepsilon_i \tau_{k_j} \leq \| x^{k_j+1} - x^{k_j} \|.$$

This in return implies

$$F(x^{k_j}) - F(x^{k_j+1}) \geq \varepsilon_i^2 \tau_{\min}.$$

On the other hand, if $\lambda \geq \delta_i$, then

$$F(x^{k_j}) - F(x^{k_j+1}) \geq \frac{\delta_i^2}{\tau_{\max}}.$$

Setting $\mu = \min\left\{\varepsilon_i^2 \tau_{\min}, \frac{\delta_i^2}{\tau_{\max}}\right\}$, it follows that whenever $x^{k_j} \in B_i$ and $d^{k_j+1} \in B_{\delta_i}(\widetilde{d^i})$, then

$$F(x^{k_j}) - F(x^{k_j+1}) \geq \mu.$$

Choosing $K \in \mathbb{N}$ such that $K\mu > M - m$, we know that this event only has to occur K times for $(x^{k_j})_{j\in\mathbb{N}}$ to leave B_i. In other words, almost surely, there is no subsequence $(x^{k_j})_{j\in\mathbb{N}} \subset B_i$. This concludes the proof. □

Deterministic case

We now cover the deterministic case, in which $(d^k)_{k\in\mathbb{N}}$ is required to be *cyclically dense*.

Definition 4.7. *A sequence $(d^k)_{k\in\mathbb{N}} \subset S^{n-1}$ is* cyclically dense *in S^{n-1} if, for all $\varepsilon > 0$, there is $N \in \mathbb{N}$ such that for any $k \in \mathbb{N}$, the set $\left\{d^k, \ldots, d^{k+N-1}\right\}$ forms an ε-cover of S^{n-1},*

$$S^{n-1} \subset \bigcup_{i=k+1}^{k+N-1} B_\varepsilon(d^i).$$

Remark 4.8. *Randomly drawn sequences are almost surely not cyclically dense, hence the separate treatment of the stochastic and deterministic methods.*

Many constructions of dense sequences are also cyclically dense. We provide an example of such a sequence on the unit interval $[0,1]$.

Example 4.9. *Let $\sigma \in (0,1)$ be an irrational number and define the sequence $(\lambda_k)_{k\in\mathbb{N}}$ in $[0,1]$ by*

$$\lambda_k = (\sigma k) \pmod 1 = \sigma k - \lfloor \sigma k \rfloor,$$

where $\lfloor \sigma k \rfloor$ denotes the largest integer less than or equal to σk.

To see that $(\lambda_k)_{k\in\mathbb{N}}$ is cyclically dense in $[0,1]$, set $\varepsilon > 0$ and note by sequential compactness of $[0,1]$ that there is $k, r \in \mathbb{N}$ such that $|\lambda_k - \lambda_{k+r}| < \varepsilon$. We can write $\delta = |\lambda_k - \lambda_{k+r}| > 0$, where we know that δ is strictly positive, as no value can be repeated in the sequence due to

σ being irrational. By modular arithmetic, we have for any $l \in \mathbb{N}$,

$$\lambda_{k+rl} = \lambda_k + l\delta \pmod{1}.$$

In other words, the subsequence $(\lambda_{k+rl})_{l\in\mathbb{N}}$ moves in increments of $\delta < \varepsilon$ on $[0,1]$. Setting $N = r\left\lceil \frac{1}{\delta} \right\rceil + k$, where $\lceil \delta \rceil$ denotes the smallest integer greater than or equal to δ, it is clear that for any $j \in \mathbb{N}$, the set $\{\lambda_j, \lambda_{j+1}, \dots, \lambda_{j+N-1}\}$ forms an ε-cover of $[0,1]$.

One could naturally extend this construction to higher dimensions $[0,1]^n$, by choosing n irrational numbers such that any (non-zero) linear combinations with rational coefficients is also irrational.

Theorem 4.10. *Let $(x^k)_{k\in\mathbb{N}}$ solve (3.4), where $(d^k)_{k\in\mathbb{N}}$ are cyclically dense. Then all accumulation points $x^* \in S$ satisfy $0 \in \partial F(x^*)$.*

Proof. We consider the setup in the proof to Theorem 4.6, where X is the set of nonstationary points (4.3) and is covered by a countable collection of open balls (4.5),

$$X \subset \bigcup_{i\in\mathbb{N}} B_{\delta_i}(y^i).$$

We will show that an accumulation point $x^* \in S$ cannot belong to the ball $B_{\delta_i}(y^i)$, from which it follows that S is a subset of the set of stationary points. For contradiction, suppose that there is a subsequence $(x^{k_j})_{j\in\mathbb{N}} \to x^* \in B_{\delta_i}(y^i)$. By Lemma 4.3 *(iii)*, since $\|x^k - x^{k+1}\| \to 0$ as $k \to \infty$, we deduce that for any $N \in \mathbb{N}$, there is $j \in \mathbb{N}$ such that

$$\{x^{k_j}, x^{k_j+1}, \dots, x^{k_j+N-1}\} \subset B_{\delta_i}(y^i).$$

Then, by cyclical density, we can choose N such that the directions $\{d^{k_j}, d^{k_j+1}, \dots, d^{k_j+N-1}\}$ form an ε_i-cover of S^{n-1}. Therefore, there exists $x^k \in B_{\delta_i}(y^i)$ and $d^k \in B_{\delta_i}(\widetilde{d^i})$, so we can argue as in Theorem 4.6, that

$$F(x^k) - F(x^{k+1}) \geq \mu,$$

where $\mu = \min\left\{\varepsilon_i^2 \tau_{\min}, \frac{\delta_i^2}{\tau_{\max}}\right\}$. If $(x^{k_j})_{j\in\mathbb{N}}$ had a limit in $B_{\delta_i}(y^i)$, this would happen arbitrarily many times, which is a contradiction. This concludes the proof. □

4.2.3 Necessity of search density and Lipschitz continuity

For nonsmooth problems, it is necessary to employ a set of directions $(d^k)_{k\in\mathbb{N}}$ larger than the set of basis coordinates $\{e^1, \dots, e^n\}$. To see this, we can consider the function $F(x,y) =$

$\max\{x,y\}$ and the starting point $x^0 = [1,1]^T$. With the standard Itoh–Abe discrete gradient method, the iterates would remain at x^0, even though this point is nonstationary.

We show with a simple example that the assumption of density of $(d^k)_{k\in\mathbb{N}}$ in Theorem 4.6 is not only sufficient, but also necessary.

Example 4.11. *We suppose $F : \mathbb{R}^2 \to \mathbb{R}$ is defined by $F(x_1,x_2) = |x_1| + N|y_2|$ for some $N \in \mathbb{N}$, and set $x^0 = [-1,0]^T$. For $\theta \in [-\pi/2,\pi/2]$, let $d = [\cos\theta,\sin\theta]^T$. Then $-d$ is a direction of descent if and only if $\theta \in (-\arctan(1/N),\arctan(1/N))$. This interval can be made arbitrarily small by choosing N to be sufficiently large. Therefore, for an Itoh–Abe method to descend from x^0 for arbitrary functions, the directions $(d^k)_{k\in\mathbb{N}}$ need to include a convergent subsequence to the direction $[1,0]^T$. As this direction is arbitrary, we deduce that $(d^k)_{k\in\mathbb{N}}$ must be dense.*

Theorem 4.6 also assumes that F is locally Lipschitz continuous. We briefly discuss why this assumption is necessary, and provide an example to show that for functions that are merely continuous, the theorem no longer holds.

By Proposition 2.1.1. (b) in [54], the mapping $(y,d) \mapsto F^o(y;d)$ is upper semicontinuous for y in a neighbourhood of x, due to the local Lipschitz continuity of F near x. That is,

$$F^o(y^*;d^*) \geq \limsup_{y\to y^*, d\to d^*} F^o(y;d).$$

This property is crucial for the convergence analysis of Itoh–Abe methods, as it implies

$$F^o(x^*,d^*) \geq \limsup_{k\in\mathbb{N}} F^o(x^k;d^k) = 0.$$

Without local Lipschitz continuity, it is possible to have

$$x^k \to x^*, \quad d^k \to d^*, \text{ and } F^o(x^k;d^k) \to 0, \qquad \text{but } F^o(x^*;d^*) < 0.$$

In this case, there is no guarantee that the limit x^* is Clarke stationary. We demonstrate this with an example.

Example 4.12. *We will first state the iterates $(x^k)_{k\in\mathbb{N}}$ and then construct a function $F : \mathbb{R}^2 \to \mathbb{R}$ that fits these iterates. Let $(d^k)_{k\in\mathbb{N}}$ be a cyclically dense sequence in S^1 and assume without loss of generality that $[0,1]^T \notin (d^k)_{k\in\mathbb{N}}$. Replacing d^k with $-d^k$ does not change the step in* (3.4)*, so we assume that $d_1^k < 0$ for all k. We set $x^0 = [0,0]^T$ and define $(x^k)_{k\in\mathbb{N}}$ and $(F(x^k))_{k\in\mathbb{N}}$ to be*

$$x^{k+1} = x^k - \frac{1}{(k+1)^2}d^k, \qquad F(x^{k+1}) = F(x^k) - \frac{1}{(k+1)^4}, \qquad F(x^0) = 0.$$

Since $\sum_{k\in\mathbb{N}} \|x^k - x^{k+1}\| < \infty$*, it follows that* x^k *converges to some limit* x^**, and* $F(x^k)$ *clearly decreases to a limit* $F^* \in \mathbb{R}$*. Furthermore, these steps satisfy* (3.4) *with* $\tau_k = 1$*. We then define* F *on the line segments* $[x^k, x^{k+1}] := \{\lambda x^k + (1-\lambda)x^{k+1} \ : \ \lambda \in [0,1]\}$ *by interpolating linearly from* $(x^k, F(x^k))$ *to* $(x^{k+1}, F(x^{k+1}))$*.*

Next, we define F *on* $\mathbb{R}^2$ *as a function that linearly decreases everywhere in the direction* $[0,1]^T$ *at the rate of 1, and so that its value is consistent with the values given on the predefined line segments* $[x^k, x^{k+1}]$*. Note that this is a well-defined and continuous function, since each line in the direction* $[0,1]^T$ *crosses at most one point on at most one line segment, due to our assumptions on* $(d^k)_{k\in\mathbb{N}}$*.*

We conclude the example by noting that the limit x^* *is not Clarke stationary—in fact, no point is Clarke stationary—since* $F^o(x;[0,1]^T) = -1$ *for all* x*.*

4.2.4 Nonsmooth, nonconvex functions with further regularity

For a large class of nonsmooth optimisation problems (convex and nonconvex), the objective function is sufficiently regular so that the standard Itoh–Abe discrete gradient method is also guaranteed to converge to Clarke stationary points. These are functions F for which $x^* \in \mathbb{R}^n$ is Clarke stationary if and only if $F^o(x^*;e^i) \geq 0$ for $i = 1, \dots, n$. One example is functions of the form

$$F(x) = E(x) + \lambda \|x\|_1,$$

where E is a continuously differentiable function that may be nonconvex, $\|x\|_1$ denotes $|x_1| + \ldots + |x_n|$, and $\lambda > 0$. See for example Proposition 2.3.3 and the subsequent corollary in [54], combined with the fact that the nonsmooth component of F, i.e. $\|\cdot\|_1$, separates into n coordinate-wise scalar functions. This implies that the Clarke subdifferential is given by

$$\partial F(x) = \{\nabla E(x)\} + \lambda \prod_{i=1}^{n} \mathrm{sgn}(x_i),$$

where $\prod$ denotes the Cartesian product and

$$\mathrm{sgn}(x_i) := \begin{cases} \{1\}, & \text{if } x_i > 0, \\ \{-1\}, & \text{if } x_i < 0, \\ [-1,1], & \text{if } x_i = 0. \end{cases}$$

Since this chapter is chiefly concerned with the blackbox setting where no particular structure of F is assumed, we do not include a rigorous analysis of the convergence properties of the standard Itoh–Abe discrete gradient method for functions of the above form. However, we point out that for nonsmooth, nonconvex optimisation problems where Clarke stationarity is equivalent to Clarke directional stationarity along the standard coordinates, one can adapt Theorem 4.6 in a straightforward manner to prove that the iterates converge to a set of Clarke stationary points when the directions $(d^k)_{k\in\mathbb{N}}$ are drawn from the standard coordinates $(e^i)_{i=1}^n$.

Furthermore, one could drop the requirement that F is locally Lipschitz continuous, and replace $\|x\|_1$ with $\|x\|_p^p$, where $p \in (0,1)$, and $\|x\|_p^p = |x_1|^p + \ldots + |x_n|^p$.

4.3 Rotated Itoh–Abe discrete gradients

We briefly discuss a generalised Itoh–Abe method that retains the Itoh–Abe discrete gradient structure, by ensuring that the directions $(d^{kn+1}, d^{kn+2}, \ldots, d^{k(n+1)})$ are orthonormal. Equivalently, we consider each block of n directions to be independently drawn from a random distribution on the set of orthogonal transformations on $\mathbb{R}^n$ with determinant 1, denoted by $\mathrm{SO}(n)$.

Definition 4.13. *The* orthogonal group of dimension *n, $SO(n)$, is the set of orthogonal matrices in $\mathbb{R}^n$ with determinant* 1*, so if $R \in SO(n)$, then $R^{-1} = R^T$. Therefore R maps one orthonormal basis of $\mathbb{R}^n$ to another.*

Each element of $\mathrm{SO}(n)$ corresponds to a *rotated* Itoh–Abe discrete gradient.

Definition 4.14 (Rotated Itoh–Abe discrete gradient)**.** *Suppose $R \in SO(n)$ maps the basis $(e^i)_{i=1}^n$ to another orthonormal basis $(f^i)_{i=1}^n$, i.e. $Rf^i = e^i$. For continuously differentiable functions F, the* rotated Itoh–Abe discrete gradient*, denoted by $\overline{\nabla}_R F$, is given by*

$$\overline{\nabla}_R F(x,y) = R^T \hat{\nabla}_R F(x,y),$$

where

$$\left(\hat{\nabla}_R F(x,y)\right)_i := \frac{F\left(x + \sum_{j=1}^{i} \langle y-x, f^j\rangle f^j\right) - F\left(x + \sum_{j=1}^{i-1} \langle y-x, f^j\rangle f^j\right)}{\langle y-x, f^i\rangle}.$$

It is straightforward to check that it is a discrete gradient, as defined for continuously differentiable functions F.

Proposition 4.15. *$\overline{\nabla}_R F$ is a discrete gradient.*

Proof. For any $x, y \in \mathbb{R}^n$, $x \neq y$,

$$\begin{aligned}
&\langle \overline{\nabla}_R F(x,y), y-x\rangle \\
&\qquad = \langle R^T \hat{\nabla}_R F(x,y), y-x\rangle \\
&\qquad = \langle \hat{\nabla}_R F(x,y), R(y-x)\rangle \\
&\qquad = \sum_{i=1}^n \frac{F\left(x+\sum_{j=1}^i \langle y-x, f^j\rangle f^j\right) - F\left(x+\sum_{j=1}^{i-1}\langle y-x, f^j\rangle f^j\right)}{\langle y-x, f^i\rangle} \cdot \langle y-x, f^i\rangle \\
&\qquad = \sum_{i=1}^n F\left(x+\sum_{j=1}^i \langle y-x, f^j\rangle f^j\right) - F\left(x+\sum_{j=1}^{i-1}\langle y-x, f^j\rangle f^j\right) \\
&\qquad = F(y) - F(x).
\end{aligned}$$

The convergence property $\lim_{y\to x} \overline{\nabla}_R F(x,y) = \nabla F(x)$ is immediate, providing F is continously differentiable. □

Thus, we can implement schemes that are formally discrete gradient methods, and also fulfill the convergence theorems in Section 4.2.

4.4 Numerical implementation

We consider three ways of choosing $(d^k)_{k\in\mathbb{N}}$.

1. *Standard Itoh–Abe method.* The directions cycle through the standard coordinates, with the rule $d^k = e^{[(k-1) \bmod n]+1}$. Performing n steps of this method is equivalent to one step with the standard Itoh–Abe discrete gradient method.

2. *Random pursuit.* The directions are independently drawn from a random distribution Ξ on S^{n-1}. We assume that the support of the density of Ξ is dense in S^{n-1}.

3. *Rotated Itoh–Abe method.* For each $k \in \mathbb{N}$, the block of n consecutive directions $(d^{kn+1}, d^{kn+2}, \ldots, d^{(k+1)n})$ is drawn from a random distribution on $\mathrm{O}(n)$, the orthogonal group of dimension n. In other words, the directions form an orthonormal basis. This retains the discrete gradient structure of the standard Itoh–Abe discrete gradient method. We assume that each draw from $\mathrm{O}(n)$ is independent, and, for notational continuity, we denote by Ξ the marginal distribution of d^{kn+1} on S^{n-1}, and again assume that the support of the density is dense in S^{n-1}.

We formalise an implementation of randomised Itoh–Abe methods with two algorithms, an inner and an outer one. Algorithm 3 is the inner algorithm and solves (3.4) for x^{k+1}, given x^k, d^k and time step bounds $\tau_{\min}, \tau_{\max}$. Algorithm 2 is the outer algorithm, which calls the inner algorithm for each iterate x^k, and provides a stopping rule for the methods. The stopping rule in Algorithm 2 takes two positive integers K and M as parameters, such that the algorithm stops either after K iterations, or when the iterates have not sufficiently decreased F in the last M iterations. We typically set $M \approx n$, n being the dimension of the domain. The exception to this is when the function F is expected to be highly irregular or nonsmooth, in which case we choose a larger M, as directions are generally prone to yield insufficient decrease. This stopping rule can be replaced by any other heuristic.

Algorithm 3 is a tailormade scalar solver for (3.4) that balances the tradeoff between optimally decreasing F given constraints $\tau_{\min}, \tau_{\max}$ and using minimal function evaluations. Rather than solving for a given τ_k, it ensures that there exists some $\tau_k \in [\tau_{\min}, \tau_{\max}]$ that matches the output x^{k+1}. It requires a preliminary $\tau \in [\tau_{\min}, \tau_{\max}]$, which we heuristically chose as $\tau = \sqrt{\tau_{\min}\tau_{\max}}$. This method is particularly suitable when $\tau_{\min} \ll \tau_{\max}$, and can be replaced by any other scalar root finder algorithm.

The generalised Itoh–Abe methods have been implemented on Python.

Algorithm 2 Generalised Itoh–Abe method with solver and stopping criterion

Input: starting point x^0, directions $(d^k)_{k\in\mathbb{N}}$, time step bounds $(\tau_{\min}, \tau_{\max})$, tolerance for function reduction η, maximal number of iterations K, maximal number of consecutive directions without descent before stopping M, internal solver described by Algorithm 3.
Initialise: counter $m = 0$.

```
for $k = 0, \dots, K-1$ do
    Update $x^{k+1} \hookleftarrow (x^k, d^{k+1}, \tau_{\min}, \tau_{\max})$ via Algorithm 3
    if $F(x^k) - F(x^{k+1}) \leq \eta$ then
        $m = m + 1$
    else
        $m = 0$
    end if
    if $m \geq M$ then
        Terminate
    end if
end for
```

4.5 Examples

In this section, we use the generalised Itoh–Abe methods to solve several nonsmooth, nonconvex problems. In Section 4.5.1, we consider some well-known optimisation challenges

Algorithm 3 Solver for Itoh–Abe step (3.4)

Input: current point x, direction d, time step upper bound $\tau_{\max}$, time step lower bound $\tau_{\min}$, predicted time step $\tau = \sqrt{\tau_{\min}\tau_{\max}}$, tolerance for x, ε, scalar $\sigma \in (0,1)$.

```
if $F(x+\varepsilon d) \geq F(x)$ then
    $d = -d$
    if $F(x+\varepsilon d) \geq F(x)$ then
        return $x$ (stationary along $d$)
    end if
end if
Solve for $\alpha$ assuming linear extrapolation of $F$ and with predicted $\tau$ (assume for simplicity $\alpha > \varepsilon$):
    $\alpha = -\frac{F(x+\varepsilon d) - F(x)}{\varepsilon\tau}$
    $x^0 = x, \quad x_1 = x + \varepsilon d, \quad x_2 = x + \alpha d$
while $F$ is concave between $x^0$, $x^1$ and $x^2$ (meaning $\frac{F(x^2)-F(x^1)}{x^2-x^1} \leq \frac{F(x^1)-F(x^0)}{x^1-x^0}$) do
    $\alpha = \alpha/\sigma, x^2 = x + \alpha d.$
end while
Do step of parabolic interpolation (see [107, Section 6.2.2]) between $x^0$, $x^1$ and $x^2$, i.e.
    $y = x^1 - \frac{1}{2}\frac{(x^1-x^0)^2(F(x^1)-F(x^2)) - (x^1-x^2)^2(F(x^1)-F(x^0))}{(x^1-x^0)(F(x^1)-F(x^2)) - (x^1-x^2)(F(x^1)-F(x^0))}$
while Parabolic step has not decreased $F$ do
    Update parabolic interpolation points $x^i$, $i = 0,1,2$.
end while
$y = x^i$ is optimal point from parabolic interpolation step
while $\frac{|F(y)-F(x)|}{\|y-x\|^2} \notin \left[1/\tau_{\max}, 1/\tau_{\min}\right]$ do
    if $\frac{|F(y)-F(x)|}{\|y-x\|^2} > 1/\tau_{\min}$ then
        $y = y/\sigma$
    else
        $y = \sigma y$
    end if
end while
return $y$
```

developed by Rosenbrock and Nesterov. In Section 4.5.2, we solve bilevel optimisation of parameters in variational regularisation problems.[3]

[3] Test images are taken from the Berkeley database [146]. Available online: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html.

We compare our method to state-of-the-art derivative-free optimisation methods Py-BOBYQA [42, 181] and the LT-MADS solver provided by NOMAD [9, 128, 127]. For purposes of comparing results across solvers for these problems, we do not measure objective function value against iterates, but objective function value against function evaluations.

4.5.1 Rosenbrock functions

We consider the well-known Rosenbrock function [194]

$$F(x,y) = (1-x)^2 + 100(y-x^2)^2. \tag{4.6}$$

Its global minimiser $[1,1]^T$ is located in a narrow, curved valley, which is challenging for the iterates to navigate. We compare the three variants of the Itoh–Abe method, for which we set the algorithm parameters $\varepsilon = 10^{-5}$, $\tau_{\min} = 10^{-4}$, $\tau_{\max} = 10^2$, $\eta = 10^{-9}$, and $M = 30$. See Figure 4.1 for the numerical results. All three methods converge to the global minimiser, which shows that the Itoh–Abe methods are robust. Unsurprisingly, the random pursuit method and the rotated Itoh–Abe method, which descend in varying directions, perform significantly better than the standard Itoh–Abe method.

We additionally consider a nonsmooth variant of (4.6), termed Nesterov's (second) nonsmooth Chebyshev–Rosenbrock function [102],

$$F(x,y) = \frac{1}{4}|x-1| + \big|y - 2|x| + 1\big|. \tag{4.7}$$

In this case too, the global minimiser $[1,1]^T$ is located along a narrow path. Furthermore, there is a nonminimising, stationary point at $[0,-1]^T$, which is nonregular—i.e. it has negative directional derivatives.

We also compare the three Itoh–Abe methods for this example, and set the algorithm parameters $\varepsilon = 10^{-10}$, $\tau_{\min} = 10^{-4}$, $\tau_{\max} = 10^2$, $\eta = 10^{-16}$, and $M = 100$. See Figure 4.2 for the results from this. As can be seen, the standard Itoh–Abe discrete gradient method is not suitable for the irregular paths and nonsmooth kinks of the objective function, and stagnates early on. The two randomised Itoh–Abe methods perform better, as they descend in varying directions. For the remaining 2D problems in this chapter, we will consider the rotated Itoh–Abe method, although we could just as well have used the random pursuit method. For higher-dimensional problems, we recommend the random pursuit method.

We compare the performance of the randomised Itoh–Abe (RIA) method to Py-BOBYQA and LT-MADS for Nesterov's nonsmooth Chebyshev–Rosenbrock function. We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-10}$, $\tau_{\min} = 10^{-4}$, $\tau_{\max} = 10^2$, $\eta = 10^{-16}$,

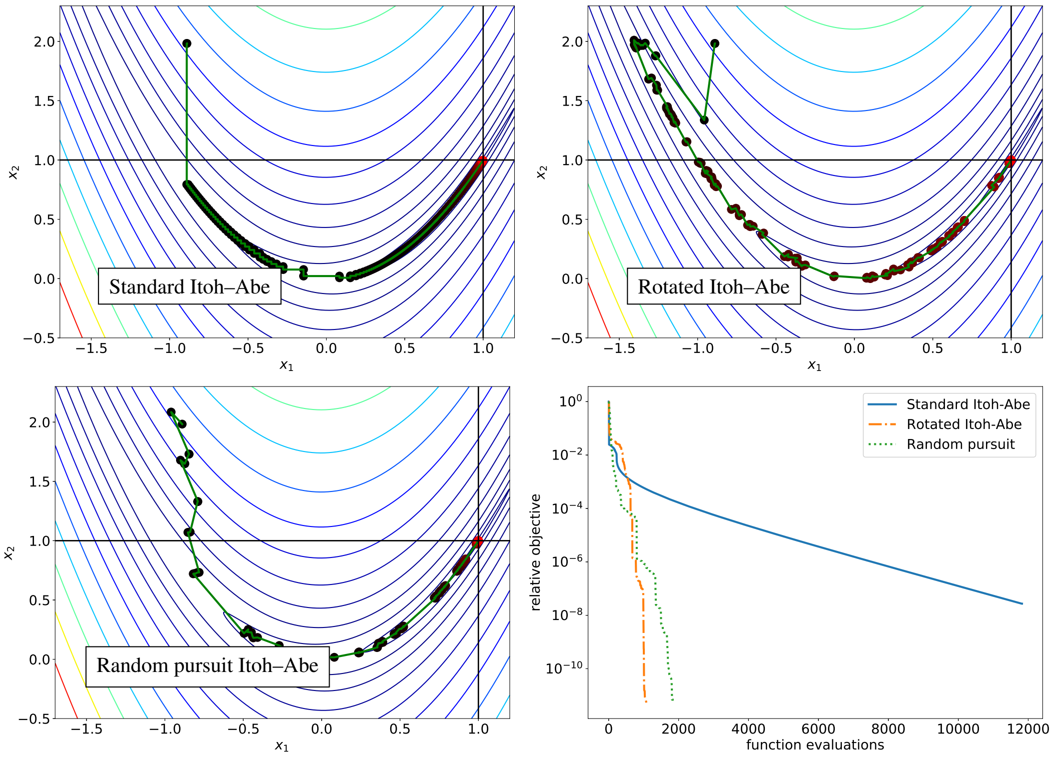


Fig. 4.1 Comparison of three variants of the Itoh–Abe method applied to the Rosenbrock function. Top left: Itoh–Abe method with standard frame. Top right: Rotated Itoh–Abe method. Bottom left: Itoh–Abe method with random pursuit. Bottom right: Convergence rates of the relative objective $\frac{F(x^k)-F^*}{F(x^0)-F^*}$ for the three variants.

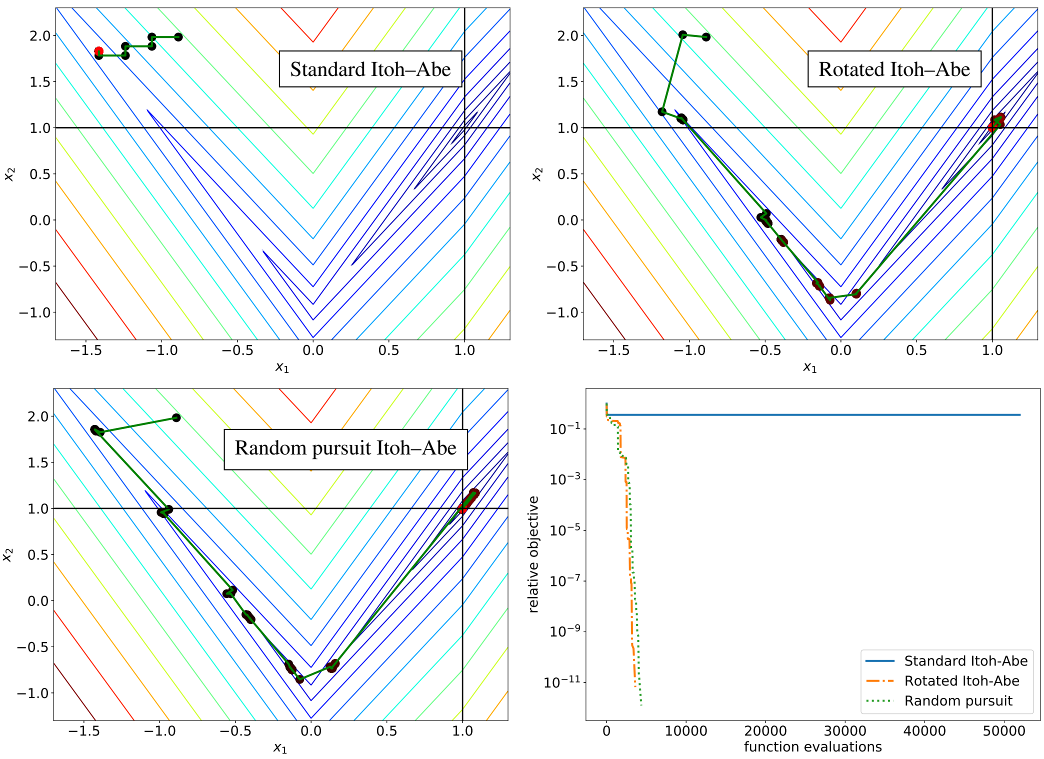


Fig. 4.2 Comparison of three variants of the Itoh–Abe method applied to Nesterov's nonsmooth Chebyshev–Rosenbrock function. Top left: Itoh–Abe method with standard frame. Top right: Rotated Itoh–Abe. Bottom left: Itoh–Abe with random pursuit. Bottom right: Convergence rates of the relative objective $\frac{F(x^k)-F^*}{F(x^0)-F^*}$ for the three variants.

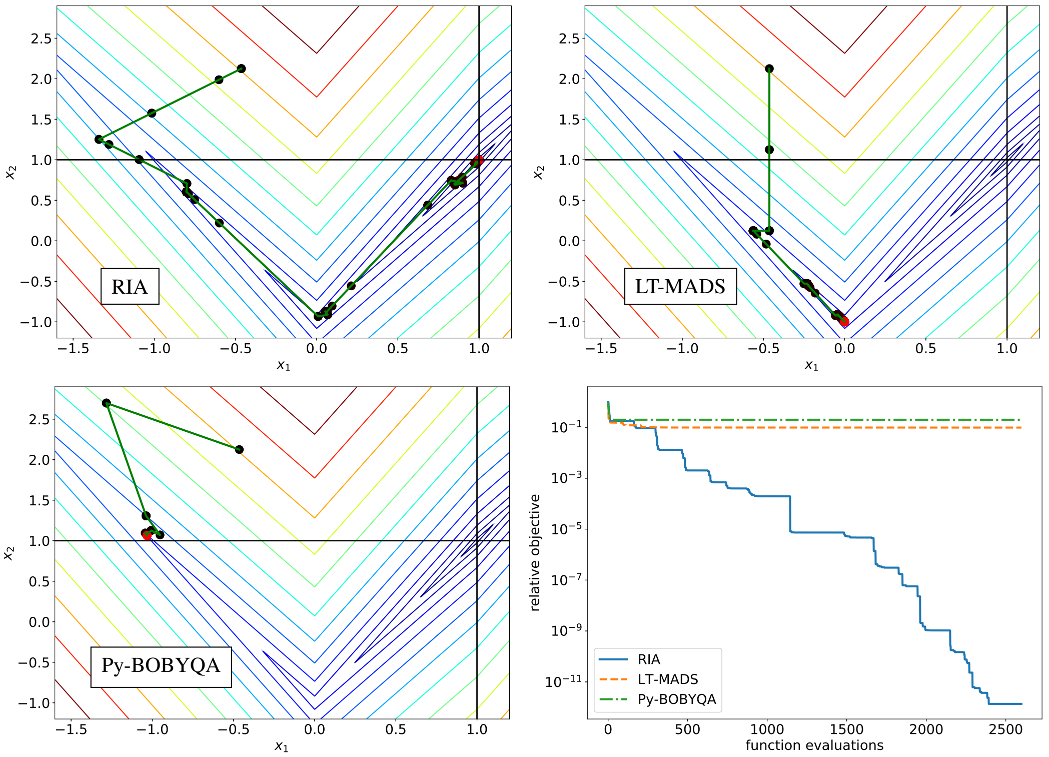


Fig. 4.3 Comparison of rotated Itoh–Abe method, LT-MADS and Py-BOBYQA applied to Nesterov's nonsmooth Chebyshev–Rosenbrock function. Top left: The iterates from the Itoh–Abe method locate the unique minimiser to an order of accuracy of about 10^{-11}. Top right: The iterates from the LT-MADS method locate the nonminimising stationary point. Bottom left: The iterates from the Py-BOBYQA method stagnate due to nonsmoothness. Bottom right: A plot of the relative objective $\frac{F(x^k)-F^*}{F(x^0)-F^*}$ with respect to function evaluations, for each method.

and $M = 100$, the parameters of Py-BOBYQA to rhobeg $= 2$, rhoend $= 10^{-16}$ and npt $= (n+1)(n+2)/2$, and the parameters of LT-MADS to DIRECTION_TYPE $=$ LT $2N$ and MIN_MESH_SIZE $= 10^{-13}$. See Figure 4.3 and 4.4 for the numerical results for two different starting points. In the first case, the Itoh–Abe method successfully converges to the global minimiser, the LT-MADS method locates the nonminimising stationary point at $[0,-1]^T$, while the Py-BOBYQA iterates stagnate at a kink, reflecting the fact that the method is not designed for nonsmooth functions. In the second case, both the Itoh–Abe method and LT-MADS locate the minimiser, while the Py-BOBYQA iterates stagnate at a kink.

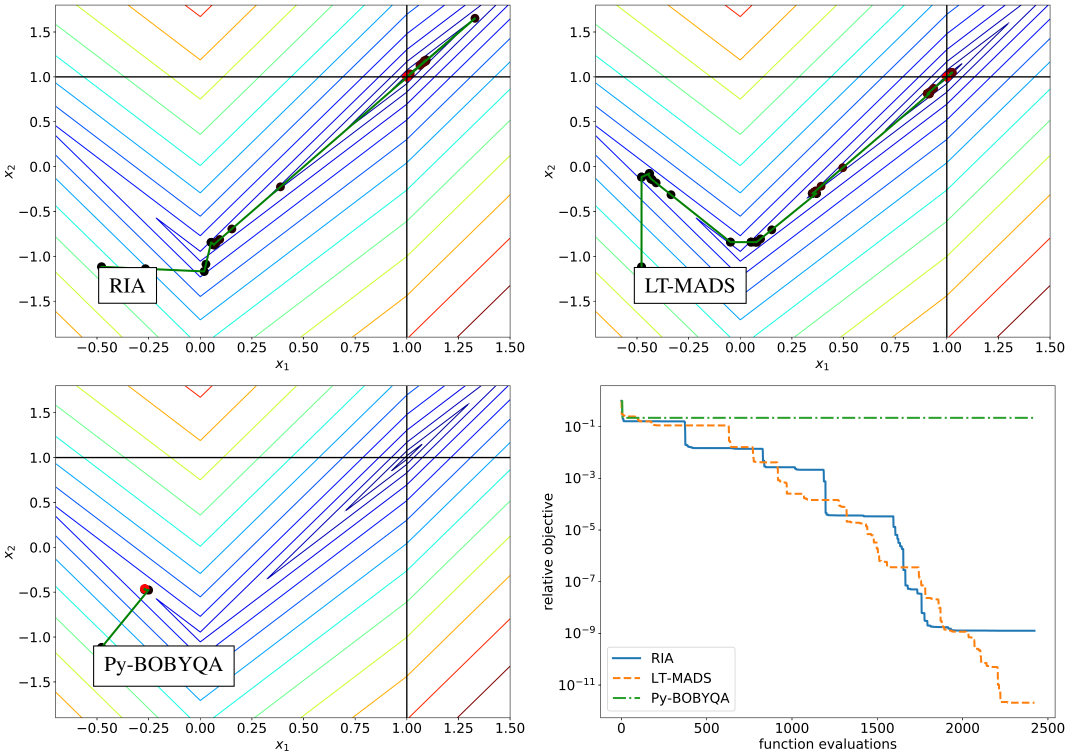


Fig. 4.4 Comparison of rotated Itoh–Abe method, LT-MADS and Py-BOBYQA applied to Nesterov's nonsmooth Chebyshev–Rosenbrock function with a different starting point. Top left: The iterates from the Itoh–Abe method locate the unique minimiser to an order of accuracy of about 10^{-11}. Top right: The iterates from the Py-BOBYQA method stagnate due to nonsmoothness. Bottom left: The iterates from the LT-MADS method locate the nonminimising stationary point. Bottom right: A plot of the relative objective $\frac{F(x^k)-F^*}{F(x^0)-F^*}$ with respect to function evaluations, for each method.

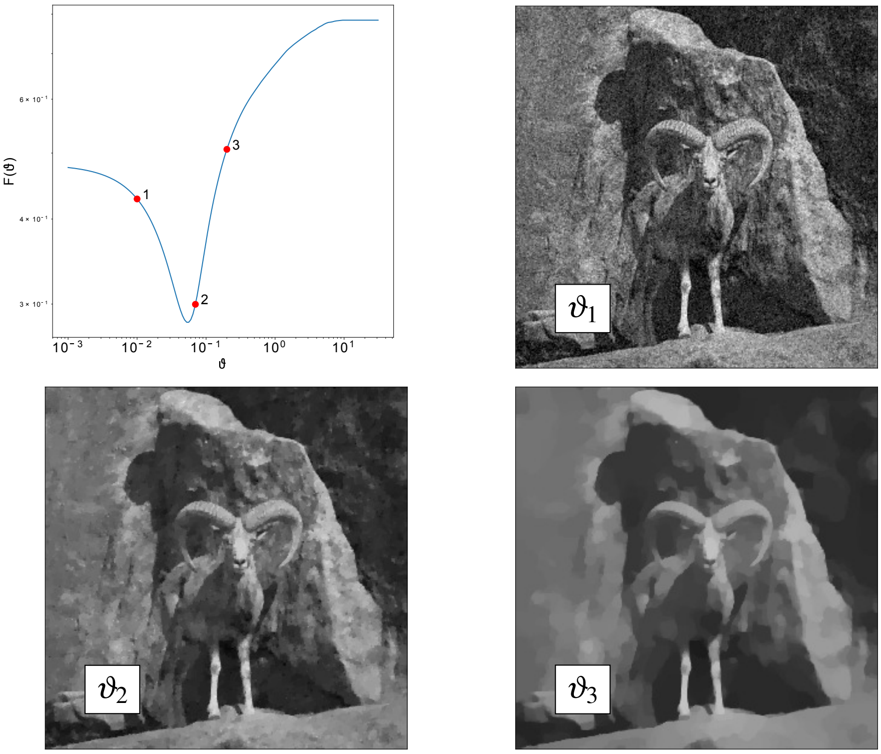


Fig. 4.5 TV denoising reconstructions for different regularisation parameters. Top left: Graph of F in (4.10). Top right: First parameter choice, $\vartheta_1 = 10^{-2}$. Bottom left: Second parameter choice, $\vartheta_2 = 7 \times 10^{-2}$. Bottom right: The third parameter choice, $\vartheta_3 = 2 \times 10^{-1}$.

4.5.2 Bilevel parameter learning in image analysis

In this subsection, we consider the Itoh–Abe method for solving bilevel optimisation problems for the learning of parameters of variational imaging problems. We restrict our focus to denoising problems, although the same method could be applied to any inverse problem. We first consider one-dimensional bilevel problems with wavelet and TV denoising, and two-dimensional problems with TGV denoising. In the TGV case, we compare the randomised Itoh–Abe method to the Py-BOBYQA and LT-MADS methods. Throughout this section, we set $M = n$, where $n = 1,2$.

Setup for variational regularisation problem

Consider an image $x^\dagger \in L^2(\Omega)$, for some domain $\Omega \subset \mathbb{R}^2$, and a noisy image

$$f^\delta = x^\dagger + \text{noise}.$$

To recover a clean image from the noisy one, we consider a parametrised family of regularisers,

$$\left\{R_\vartheta : L^2(\Omega) \to [0,\infty] \;:\; \vartheta \in [0,\infty)^n\right\},$$

and solve the variational regularisation problem

$$x_\vartheta \in \arg\min_x \frac{1}{2}\|x - f^\delta\|^2 + R_\vartheta(x). \tag{4.8}$$

We list some common regularisers in image analysis. *Total variation* (TV) [36, 196] is given by the function $R_\vartheta(x) := \vartheta\,\mathrm{TV}(x)$, where $\vartheta \in [0,\infty)$, and

$$\mathrm{TV}(x) := \sup\left\{\int_\Omega x(y)\,\mathrm{div}\,\phi(y)\,\mathrm{d}y \;:\; \phi \in C_c^1(\Omega;\mathbb{R}^d), \|\phi\|_\infty \leq 1\right\}.$$

This is one of the most common regularisers for image denoising. See Figure 4.5 for an example of denoising with TV regularisation. We also consider its second-order generalisation, *total generalised variation* [29, 28], $R_\vartheta(x) = \mathrm{TGV}^2_\vartheta(x)$, where $\vartheta = [\vartheta_1, \vartheta_2]^T \in [0,\infty)^2$ and

$$\begin{aligned}&\mathrm{TGV}^2_\vartheta(x)\\ &:= \sup\left\{\int_\Omega x(y)\,\mathrm{div}^2\,\phi(y)\,\mathrm{d}y \;:\; \phi \in C_c^2(\Omega;\mathrm{Sym}^2(\mathbb{R}^d)), \|\,\mathrm{div}^l\,\phi\|_\infty \leq \vartheta_{l+1},\; l = 0,1\right\}.\end{aligned}$$

Recall that we defined the discrete variants of TV and TGV in Chapter 1.

Recall that for a linear operator W on $L^2(\Omega)$, the basis pursuit regulariser

$$R_\vartheta(x) := \vartheta\|Wx\|_1$$

promotes sparsity of the image x in the dictionary of W.

As illustrated in Figure 4.5, the quality of the reconstruction is sensitive to ϑ. If ϑ is too low, the reconstruction is too noisy, while if ϑ is too high, too much detail is removed. As it is generally not possible to ascertain the optimal choice of ϑ a priori, a significant amount of time and effort is spent on parameter tuning. It is therefore of interest to improve our understanding of optimal parameter choices. One approach is to learn suitable parameters from training data. This requires a desired reconstruction $x^\dagger$, noisy data f^δ, and a scoring function $\Phi : L^2(\Omega) \to \mathbb{R}$ that measures the error between $x^\dagger$ and the reconstruction x_ϑ. The bilevel optimisation problem is given by

$$\vartheta^* \in \arg\min_{\vartheta\in[0,\infty)^n} \Phi(x_\vartheta), \qquad \text{s.t. } x_\vartheta \text{ solves (4.8).} \tag{4.9}$$

In our case, we have strong convexity in the data fidelity term, which implies that x_ϑ is unique for each $\vartheta \in [0,\infty)^n$. We can therefore define a mapping

$$F(\vartheta) := \Phi(x_\vartheta). \tag{4.10}$$

The bilevel problem (4.9) is difficult to tackle, both analytically as well as numerically. In most cases, the lower level problem (4.8) does not have a closed form formulation. Instead, a reconstruction x_ϑ is approximated numerically with an algorithm.

For the numerical experiments in this chapter, we reparametrise $F(\vartheta)$ as $F(\exp(\vartheta))$, where the exponential operator is applied elementwise on the parameters. There are two reasons for doing so. The first reason is that this chapter is concerned with unconstrained optimisation, and this parametrisation allows us to optimise on $\mathbb{R}^n$ instead of $[0,\infty)^n$. The second reason is that $\exp(\vartheta)$ has been found to be a preferable scaling for purposes of numerical optimisation. Note that in Chapter 5 we extend the Itoh–Abe optimisation framework to nonsmooth, nonconvex, *constrained* optimisation problems.

Wavelet denoising

We consider the wavelet denoising problem

$$x_\vartheta = \operatorname*{arg\,min}_{x \in L^2(\Omega)} \frac{1}{2}\|x - f^\delta\|^2 + \vartheta\|Wx\|_1,$$

where W is a wavelet transform. In particular, W is an orthonormal basis, which implies that the regularisation problem has the unique solution

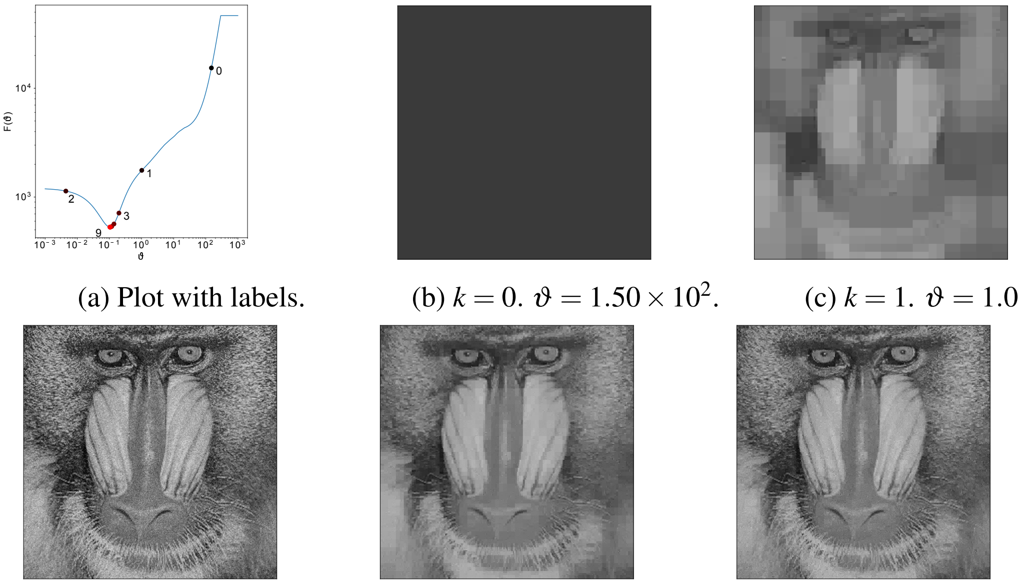
$$x_\vartheta = W^{-1}S(Wf^\delta, \vartheta),$$

where S is the shrinkage operator given in (1.5).

We first optimise ϑ for the scoring function

$$\Phi(x) := \frac{1}{2}\|x - x^\dagger\|^2.$$

We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-4}$, $\tau_{\min} = 10^{-1}$, $\tau_{\max} = 10$, and $\eta = 10^{-1}$. See Figure 4.6 for the numerical results.



(a) Plot with labels. (b) $k = 0$. $\vartheta = 1.50 \times 10^2$. (c) $k = 1$. $\vartheta = 1.02$.

(d) $k = 2$. $\vartheta = 4.42 \times 10^{-3}$. (e) $k = 3$. $\vartheta = 1.99 \times 10^{-1}$. (f) $k = 9$. $\vartheta = 1.04 \times 10^{-1}$.

Fig. 4.6 Wavelet denoising with L^2 scoring function and the Itoh–Abe method. Top left: Plot of iterates of the Itoh–Abe method. The rest: Image denoising results at different iterates k.

We also optimise ϑ with respect to the scoring function $\Phi(x) := 1 - \mathrm{SSIM}(x, x^\dagger)$, where SSIM is the *structural similarity* function [218]

$$\mathrm{SSIM}(x,y) := \frac{(2\mu_x\mu_y + c)(2\sigma_{xy} + C)}{(\mu_x^2 + \mu_y^2 + c)(\sigma_x^2 + \sigma_y^2 + C)}.$$

Here μ_x is the mean intensity of x, σ_x is the unbiased estimate of the standard deviation of x, and σ_{xy} is the correlation coefficient between x and y:

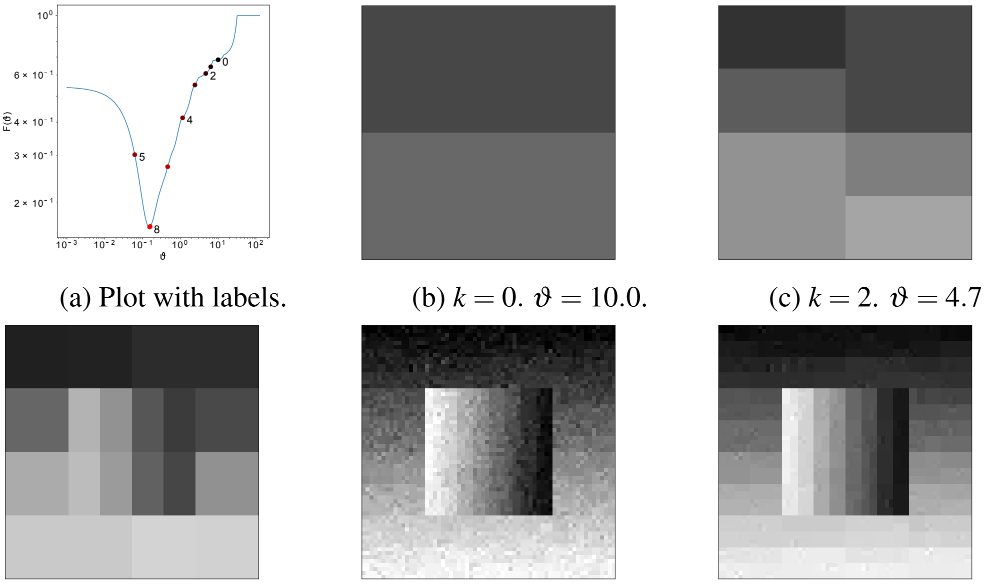
$$\mu_x := \frac{1}{m}\sum_{i=1}^{m} x_i, \;\; \sigma_x := \left(\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \mu_x)^2\right)^{\frac{1}{2}}, \;\; \sigma_{xy} := \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \mu_x)(y_i - \mu_y).$$

We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-4}$, $\tau_{\min} = 10^{-3}$, $\tau_{\max} = 10^3$, and $\eta = 10^{-2}$. See Figure 4.7 for the numerical results.

Total variation denoising

We consider the TV denoising problem

$$x_\vartheta = \operatorname*{arg\,min}_{x \in L^2(\Omega)} \frac{1}{2}\|x - f^\delta\|^2 + \vartheta\,\mathrm{TV}(x),$$



(a) Plot with labels.

(b) $k = 0$. $\vartheta = 10.0$.

(c) $k = 2$. $\vartheta = 4.71$.

(d) $k = 4$. $\vartheta = 1.15$.

(e) $k = 5$. $\vartheta = 6.23 \times 10^{-2}$.

(f) $k = 8$. $\vartheta = 1.54 \times 10^{-1}$.

Fig. 4.7 Wavelet denoising with SSIM scoring function and the Itoh–Abe method. Top left: Plot of iterates of the Itoh–Abe method. The rest: Image denoising result at different iterates k.

with the SSIM scoring function. We solve the above denoising problem using the PDHG method [50]. We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-4}$, $\tau_{\min} = 10^{-5}$, $\tau_{\max} = 9 \times 10^{-4}$, and $\eta = 10^{-5}$. See Figure 4.8 for the numerical results.

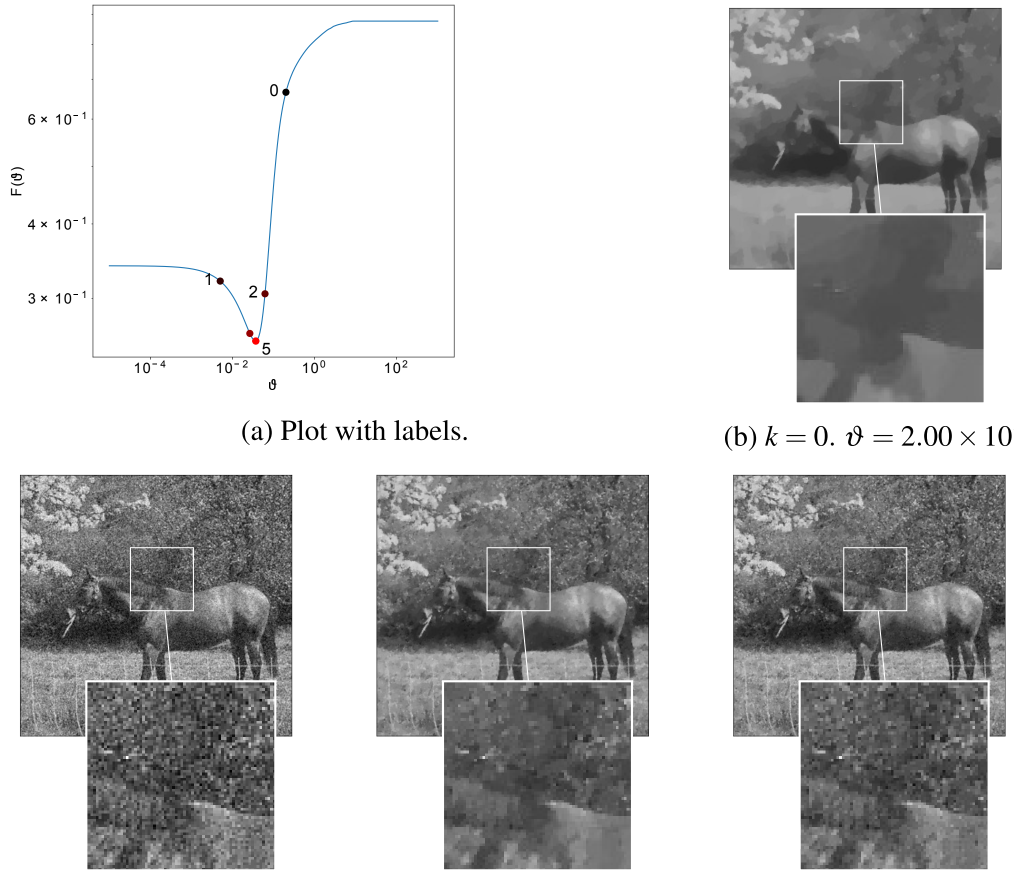
Total generalised variation regularisation

We now consider the second-order total generalised variation (TGV) regulariser for denoising, $R_{\vartheta_1,\vartheta_2}(x) = \mathrm{TGV}^2_{\vartheta_1,\vartheta_2}(x)$, with the scoring function

$$\Phi(x) := 1 - \mathrm{SSIM}(x, x^\dagger).$$

Like for TV denoising, we solve the denoising problem using the PDHG method. We set the parameters of the randomised Itoh–Abe (RIA) method to $\varepsilon = 10^{-1}$, $\tau_{\min} = 10^{-3}$, $\tau_{\max} = 10^5$, and $\eta = 10^{-20}$. See Figure 4.9 for the numerical results.

We compare these results to the results from the Py-BOBYQA and LT-MADS solvers. We set the parameters of Py-BOBYQA to rhobeg $= 2$, rhoend $= 10^{-10}$ and npt $= 2(n+1)$ and the parameters of LT-MADS to DIRECTION_TYPE = LT $2N$. See the results for two different starting points in Figure 4.10 and 4.11. We note that the objective function is approximately stationary across a range of values, which leads to the different points of



(a) Plot with labels.

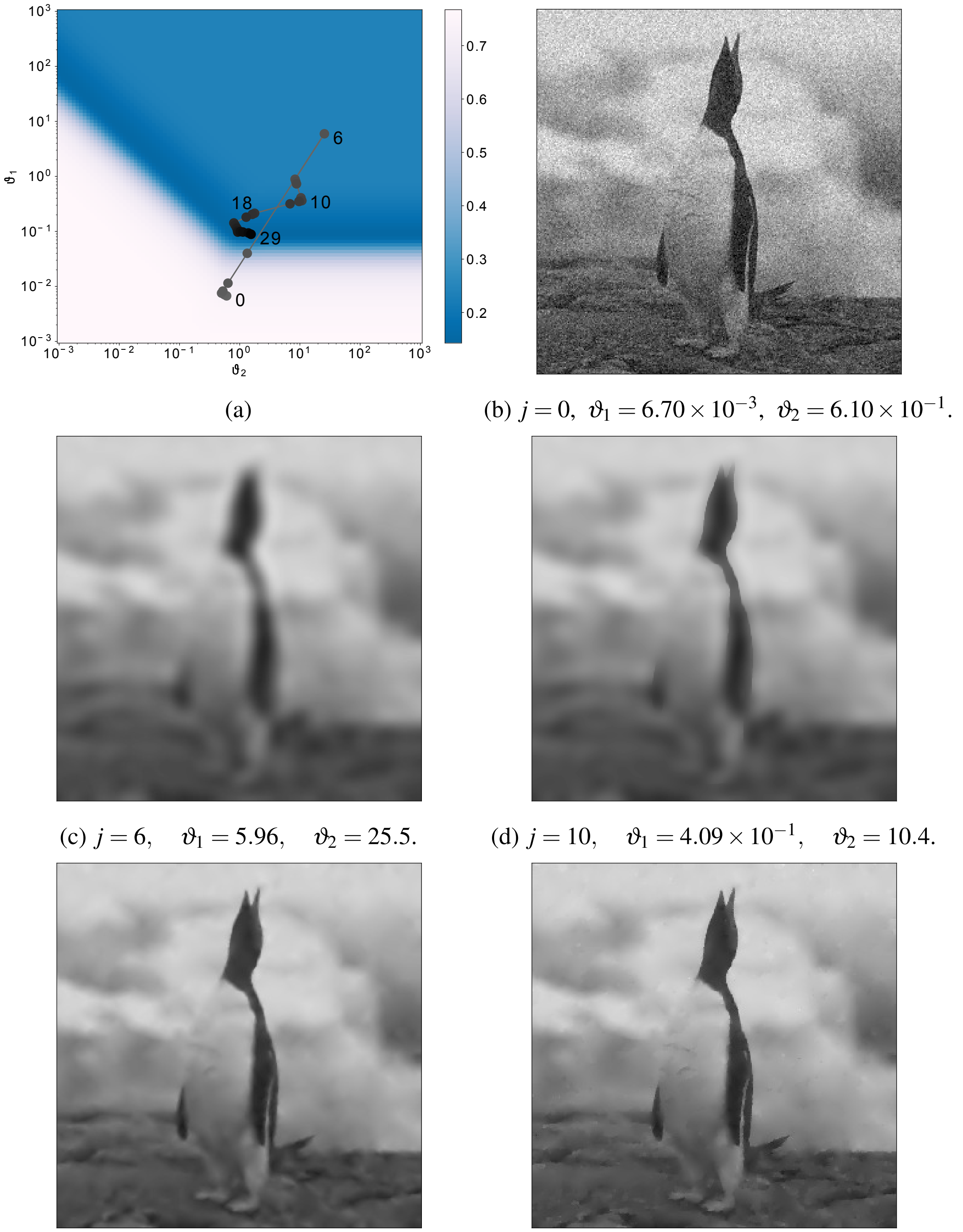
(b) $k = 0$. $\vartheta = 2.00 \times 10^{-1}$.

(c) $k = 1$. $\vartheta = 3.71 \times 10^{-3}$.

(d) $k = 2$. $\vartheta = 4.12 \times 10^{-2}$.

(e) $k = 5$. $\vartheta = 2.31 \times 10^{-2}$.

Fig. 4.8 TV denoising with SSIM scoring function and the Itoh–Abe method. Top left: Plot of iterates of the Itoh–Abe method. The rest: Image denoising result at different iterates k, with a zoom to show the difference.



(a)

(b) $j = 0,\ \vartheta_1 = 6.70 \times 10^{-3},\ \vartheta_2 = 6.10 \times 10^{-1}$.

(c) $j = 6, \quad \vartheta_1 = 5.96, \quad \vartheta_2 = 25.5$.

(d) $j = 10, \quad \vartheta_1 = 4.09 \times 10^{-1}, \quad \vartheta_2 = 10.4$.

(e) $j = 18,\ \vartheta_1 = 1.43 \times 10^{-1},\ \vartheta_2 = 7.99 \times 10^{-1}$.

(f) $j = 29,\ \vartheta_1 = 8.87 \times 10^{-2},\ \vartheta_2 = 1.55$.

Fig. 4.9 TGV denoising with SSIM scoring function and the Itoh–Abe method. Top left: Plot of iterates of the method. The rest: Image denoising result at different function evaluations j.

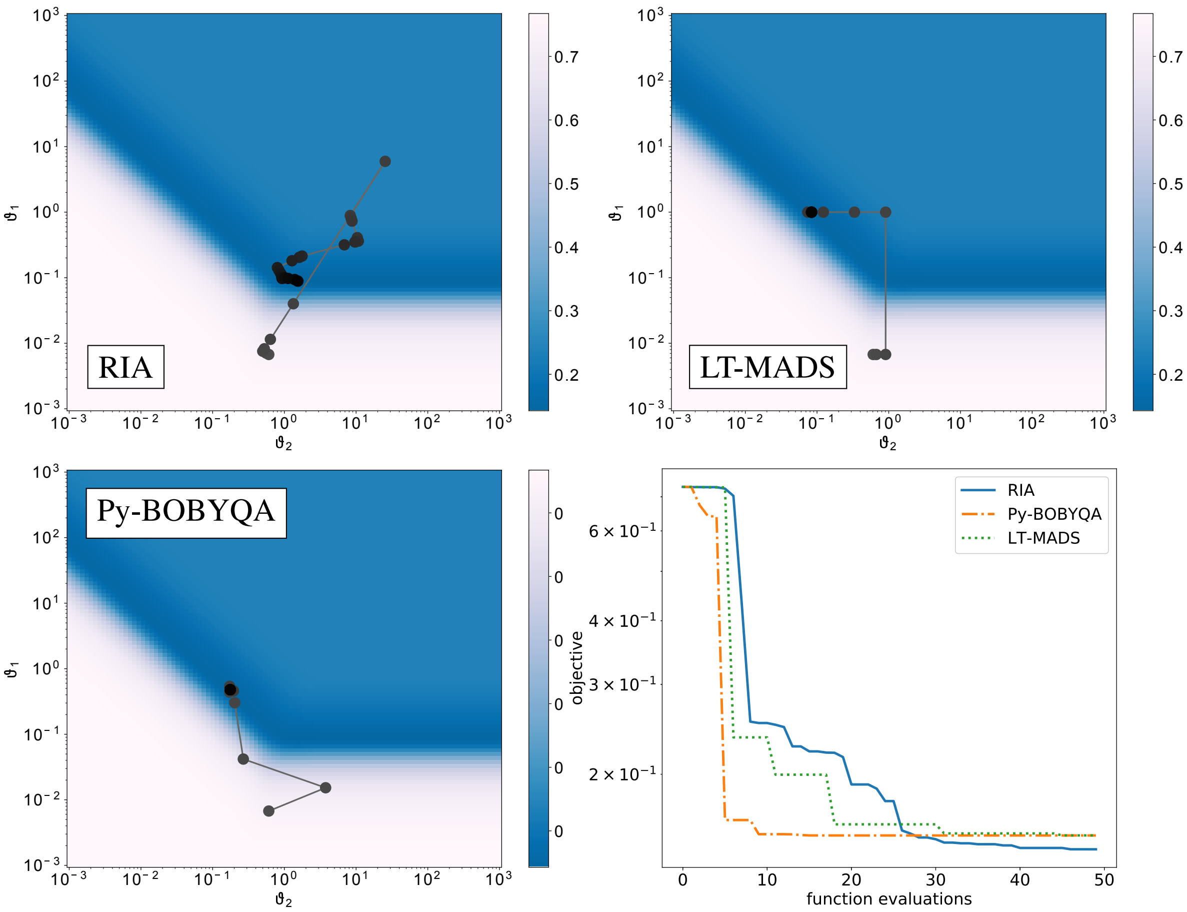


Fig. 4.10 Comparison of optimisation methods for TGV denoising with SSIM scoring function. Top left: Plot of iterates of the Itoh–Abe method. Top right: Plot of iterates of the LT-MADS method. Bottom left: Plot of iterates of the Py-BOBYQA method. Bottom right: Comparison of convergence rates for the methods with respect to function evaluations.

convergence, and different limiting values of the objective function for different methods. We see that the methods are all of comparable efficiency, although the Itoh–Abe method is slower initially. The Itoh–Abe method seems to be the most efficient, once it is within a neighbourhood of the minimiser.

4.6 Conclusion and outlook

In this chapter, we have shown that Itoh–Abe type methods are efficient and robust schemes for solving unconstrained nonsmooth, nonconvex problems without the use of gradients or subgradients. Furthermore, the favourable rates of dissipativity that the discrete gradient method inherits from the gradient flow system extends to the nonsmooth case. We show,

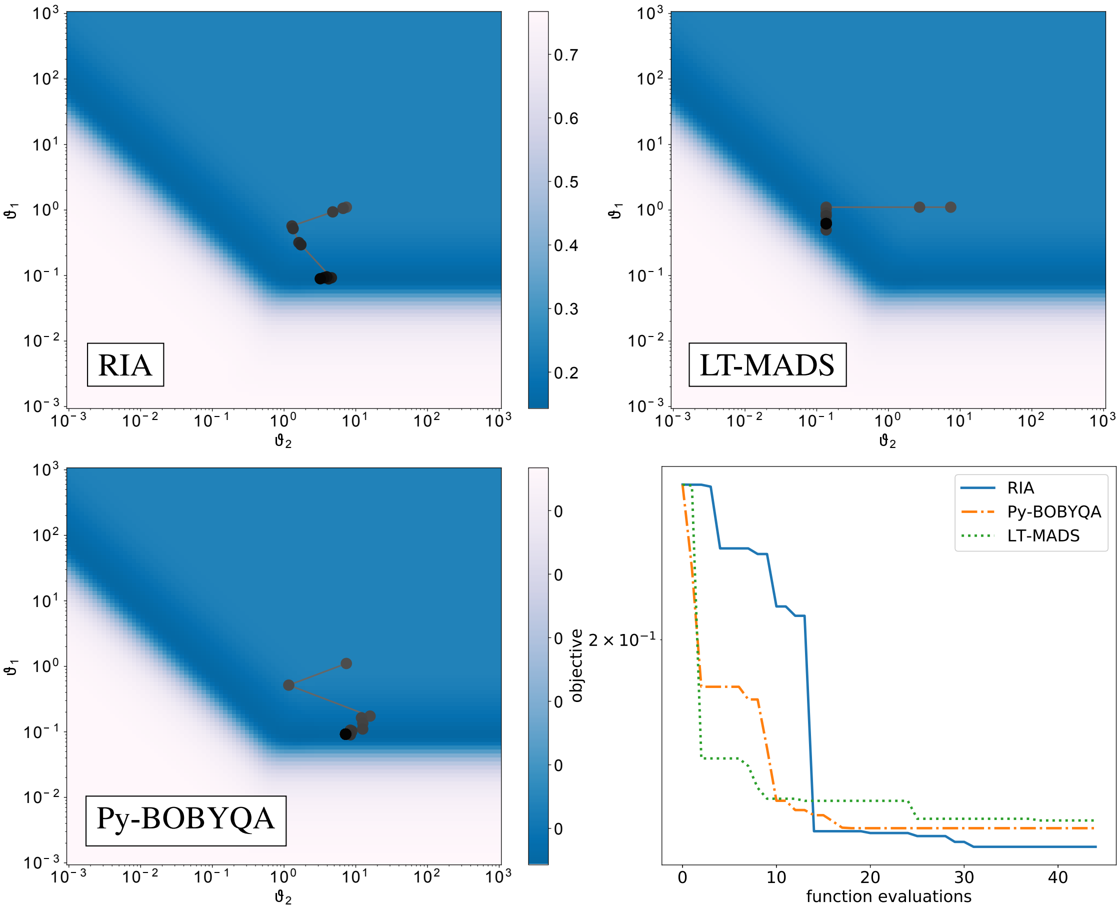


Fig. 4.11 Comparison of optimisation methods for TGV denoising with SSIM scoring function for a different starting point. Top left: Plot of iterates of the Itoh–Abe method. Top right: Plot of iterates of the LT-MADS method. Bottom left: Plot of iterates of the Py-BOBYQA method. Bottom right: Comparison of convergence rates for the methods with respect to function evaluations.

under minimal assumptions on the objective function, that the methods admit a solution that is computationally tractable, and the iterates converge to a connected set of Clarke stationary points. Through examples, the assumptions are also shown to be necessary.

The methods are shown to be versatile optimisation schemes. It locates the global minimisers of the Rosenbrock function and a variant of Nesterov's nonsmooth Chebyshev–Rosenbrock functions. The efficiency of the Itoh–Abe discrete gradient method for smooth problems has already been demonstrated elsewhere [100, 153, 185]. We also consider its application to bilevel learning problems (1.11) and compare its performance to the derivative-free Py-BOBYQA and LT-MADS methods.

Chapter 5

Discrete gradient methods for nonsmooth, nonconvex, constrained optimisation

5.1 Introduction

In Chapter 4, we looked at Itoh–Abe type methods for derivative-free optimisation of nonsmooth, nonconvex problems in the unconstrained setting. However, in the black-box setting where one is likely to consider such derivative-free methods, the problem is often subject to constraints.

In this section, we therefore consider nonsmooth, nonconvex optimisation problems with nonsmooth, nonconvex constraints, and propose a modification of the Itoh–Abe method which has convergence guarantees in this setting.

In the spirit of the previous chapter, where we sought to prove convergence assuming as little structure as possible of the objective function, we now seek to consider constraints with as little structure as possible. On that note, we consider the problem

$$\arg\min_{x\in\Omega} F(x), \tag{5.1}$$

where F is locally Lipschitz continuous, coercive and bounded below, and $\Omega \subset \mathbb{R}^n$ is an *epi-Lipschitzian set*, to be defined in the next section. Furthermore, we assume that the constraints are black-box in the sense that constraint projection maps are unavailable and that constraint feasibility can only be verified on a point-by-point basis.

Given $x \in \Omega$, a direction $d \in S^{n-1}$, and a function tolerance $\varepsilon_F \geq 0$, we want $\alpha \neq 0$ that solves

$$\left| \alpha + \tau \frac{F(x-\alpha d) - F(x)}{\alpha} \right| \leq \varepsilon_F, \quad x - \alpha d \in \Omega \tag{5.2}$$

for a time step $\tau > 0$. While in the unconstrained case, we showed that such an update exists for all time steps, in the constrained case this is no longer the case. We therefore require that each scalar update either solves (5.2) for τ, or is sufficiently close to $\operatorname{bd}\Omega$ and decreases the objective function F by any amount.

Thus we propose the following modification of the Itoh–Abe type methods in Chapter 4. For each direction $d^{k+1} \in S^{n-1}$, we assume without loss of generality that $F^o(x^k; -d^{k+1}) \leq F^o(x^k; d^{k+1})$.

Algorithm 4 Itoh–Abe method for constrained optimisation

Input: starting point $x^0 \in \mathbb{R}^n$, time steps $(\tau_k)_{k\in\mathbb{N}} \subset [\tau_{\min}, \tau_{\max}]$, directions $(d^k)_{k\in\mathbb{N}} \subset S^{n-1}$, progress parameter $\gamma \in (0,1)$, point tolerance $\varepsilon_x \geq 0$, function tolerance $\varepsilon_F \geq 0$, iteration count $k = 0$, stopping rule.

```
while not stopping rule do
    if $F(x^k - \varepsilon_x d^{k+1}) - F(x^k) > -\varepsilon_x^2/\tau_k$ (stationarity) or $x^k - \varepsilon_x d^{k+1} \notin \Omega$ (activity) then
        $x^{k+1} = x^k$.
    else
        if $\exists \lambda > \varepsilon_\Omega$ s.t. $x^k - \lambda d^{k+1} \notin \Omega$ (potentially active constraint) then
            $x^{k+1} = x^k - \alpha_k d^{k+1}$ where $\alpha_k$ either solves (5.2) for $\tau_k$, or solves (5.2) for $\widetilde{\tau}_k \in (0, \tau_k]$ such that $\alpha_k \geq \gamma\lambda$.
        else
            $x^{k+1} = x^k - \alpha_k d^{k+1}$ where $\alpha_k$ solves (5.2) for $\tau_k$.
        end if
    end if
    $k \leftarrow k + 1$
end while
```

Remark 5.1. *The intuition for the progress parameter γ is that the iterates shall always solve a standard Itoh-Abe discrete gradient descent step when possible, but otherwise, the next iterate will decrease the objective function value and progress towards the boundary at some threshold rate γ. Clearly, when $\Omega = \mathbb{R}^n$, the scheme reduces to the standard, unconstrained Itoh-Abe discrete gradient method.*

We allow for $\varepsilon_x, \varepsilon_F = 0$, accounting for cases where we can compute $F^o(x^k; -d^{k+1})$, solve the scalar update (5.2) *exactly, and check if $x^k \in \operatorname{bd}\Omega$.*

In Section 5.3, we prove that this method is implementable in a derivative-free setting, and that it comes with convergence guarantees.

5.1.1 Literature review

In this chapter, we consider the Clarke subdifferential framework for functions defined on a set $\Omega \subset \mathbb{R}^n$. For an introduction to Clarke subdifferential analysis in this setting as it relates to the *Clarke tangent cone*, see [118]. We also point out that Audet & Dennis Jr [9] applied this framework to obtain optimality guarantees for MADS.

There is a range of types of constraints for bilevel problems and more generally simulation-based optimisation, and we refer to [71] for a comprehensive classification of such constraints. Furthermore, for constrained, nonconvex optimisation when the evaluation of constraints is inexpensive compared to the evaluation of the objective function, see [43].

5.1.2 Contributions and outline

The rest of the chapter is structured as follows. In Section 5.2, we introduce epi-Lipschitzian sets, and review the theory of Clarke subdifferentials and tangent cones, in order to extend Clarke stationary points to constrained domains. In Section 5.3, we prove that the Itoh-Abe discrete gradient method for constrained problems is well-defined and computationally tractable. Furthermore, based on the theory from the previous section, we show that the iterates of the method converge to a set of constrained Clarke stationary points. In Section 5.4 we present numerical results, and in Section 5.5 we conclude.

5.2 The Clarke subdifferential and tangent cones

In this section, we review and derive properties relating to epi-Lipschitzian sets, tangent cones, and the Clarke subdifferential framework for constrained optimisation problems.

5.2.1 Epi-Lipschitzian sets

Epi-Lipschitzian sets were introduced by Rockafellar in 1978, in order to characterise sufficient and necessary regularity at the set boundary for the tangent cone to have nonempty interior [188, 190].

Definition 5.2 (Epi-Lipschitzian set)**.** *Let Ω be a subset of $\mathbb{R}^n$ and $x \in \Omega$. The set Ω is* epi-Lipschitzian *at x if there is a neighbourhood N_x of x, an invertible, linear map $A : \mathbb{R}^n \to \mathbb{R}^{n-1} \times \mathbb{R}$ and a function $\phi : \mathbb{R}^{n-1} \to \mathbb{R}$ that is locally Lipschitz continuous near the $\mathbb{R}^{n-1}$ component of $A(x)$, such that*

$$\Omega \cap N_x = \Omega \cap A^{-1}(\operatorname{epi}\phi),$$

where $\operatorname{epi}\phi := \{(\xi,t) \,:\, \phi(\xi) \le t\}$.

Ω *is* epi-Lipschitzian *if this holds for all* $x \in \Omega$.

The following property [61, Theorem 1] provides an alternative characterisation of epi-Lipschitzian sets.

Proposition 5.3. *A set* $\Omega \subset \mathbb{R}^n$ *is epi-Lipschitzian if and only if there is a locally Lipschitz continuous function* $\phi : \mathbb{R}^n \to \mathbb{R}$ *such that* Ω *and* ϕ *satisfy*

$$\begin{aligned} \Omega &= \{x \in \mathbb{R}^n \,:\, \phi(x) \le 0\}, \\ 0 &\notin \partial\phi(x) \quad \text{if } \phi(x) = 0. \end{aligned} \tag{5.3}$$

That is, Ω *is the level set of a locally Lipschitz continuous function that is not stationary on* $\operatorname{bd}\Omega$.

We recall from Chapter 2 that the Clarke tangent cone of Ω at x is given by

$$T_\Omega^{\mathrm{Cl}}(x) := \left\{ d \in \mathbb{R}^n \,:\, \exists \lambda_k \downarrow 0,\ x^k \in \Omega,\ x^k \to x,\ d^k \to d,\ \text{s.t. } x^k + \lambda_k d^k \in \Omega\ \forall k \right\}$$

We furthermore define the *hypertangent cone*.

Definition 5.4. *The* hypertangent cone $T_\Omega^H(x)$ *consists of all* $d \in \mathbb{R}^n$ *for which there exists* $\varepsilon > 0$ *such that*

$$y + \lambda e \in \Omega \quad \text{for all} \quad y \in \Omega \cap B_\varepsilon(x), \quad e \in B_\varepsilon(d), \quad \lambda \in (0,\varepsilon). \tag{5.4}$$

The following key result summarises the relationship between the hypertangent cone and the Clarke tangent cone [54, Corollary 1 to Theorem 2.5.8], as well as a more illuminating characterisation of epi-Lipschitzian sets [188, Theorem 3].

Proposition 5.5. *Suppose* $\Omega \subset \mathbb{R}^n$ *is locally closed at* $x \in \Omega$*. Then* $T_\Omega^H(x) = \operatorname{int} T_\Omega^{\mathrm{Cl}}(x)$*.*

Furthermore, $T_\Omega^H(x) \neq \emptyset$ *if and only if* Ω *is epi-Lipschitzian at* x*.*

As Proposition 5.5 shows, a defining criteria of epi-Lipschitzian sets is that the set of tangent cones must have nonempty interior. Thus, while this class include a variety of nonsmooth, nonconvex sets, it does not include lower-dimensional objects, such as hyperplanes. Another example of a non-epi-Lipschitzian set C is the epigraph of $\sqrt{|\cdot|} : \mathbb{R} \to \mathbb{R}$, at $x = [0,0]^T$. To see this, one may verify either that $T_C^H([0,0])^T = \emptyset$ or that $\sqrt{|\cdot|}$ is not locally Lipschitz continuous near 0.

We now provide an example of a group of sets which are epi-Lipschitzian. For this we define *starshaped* sets.

Definition 5.6. *The set* Ω *is* starshaped with respect to $x^* \in \Omega$ *if, for all* $x \in \Omega$ *and all* $\lambda \in [0,1]$*, we have* $\lambda x^* + (1-\lambda)x \in \Omega$.

Example 5.7. *Suppose* Ω *is a closed set for which there is* $x \in \Omega$ *and* $\varepsilon > 0$ *such that* Ω *is starshaped with respect to all* $y \in B_\varepsilon(x)$*. Then* Ω *is epi-Lipschitzian. To see this, note that for each* $z \in \Omega$*,* $T^H_\Omega(z) \supset \{B_\varepsilon(x) - z\}$.

Remark 5.8. *This includes all convex sets with nonempty interior.*

5.2.2 Clarke stationarity for constrained problems

For nonsmooth, nonconvex, constrained optimisation problems, the following constrained adaptation of the Clarke directional derivative for domain constraints was proposed by Jahn [118].

Definition 5.9 (Constrained Clarke directional derivative)**.** *For* $x \in \Omega$ *and* $d \in \mathbb{R}^n$*, the* Clarke directional derivative constrained to Ω *is given by*

$$F^o(x;d) := \limsup_{\substack{y \to x,\, \lambda \downarrow 0 \\ y \in \Omega,\, y+\lambda d \in \Omega}} \frac{F(y+\lambda d) - F(y)}{\lambda}. \tag{5.5}$$

We derive some basic properties of this mapping.

Proposition 5.10. *The Clarke directional derivative for* F *constrained to* Ω *has the following properties.*

(i) If Ω *is epi-Lipschitzian at* x*, then* $F^o(x;d)$ *exists for all* $d \in \mathbb{R}^n$.

(ii) If F *is Lipschitz continuous near* x *with Lipschitz constant* $L > 0$*, then* $|F^o(x;d)| \le L\|d\|$ *for all* $d \in \mathbb{R}^n$.

(iii) If $x \in \operatorname{int}\Omega$*, then* $F^o(x;d)$ *reduces to the standard Clarke directional derivative for unconstrained functions.*

Proof. Property *(i)*. Let $d \in \mathbb{R}^n$. We want to show that for all $\delta > 0$, there is $y \in B_\delta(x)$ and $\lambda \in (0,\delta)$ such that $y + \lambda d \in \Omega$. By Proposition 5.5, there is $e \in T^H_\Omega(x)$ and $\varepsilon > 0$ such that for all $y \in \Omega \cap B_\varepsilon(x)$, $\eta \in (0,\varepsilon)$, and $f \in B_\varepsilon(e)$, one has $y + \eta f \in \Omega$. Choose η and λ sufficiently small such that

$$y + \eta e \in B_\delta(x), \qquad e + \frac{\lambda}{\eta} d \in B_\varepsilon(e).$$

It follows that $y+\eta e+\lambda d\in\Omega$. Therefore the limit in (5.5) is well-defined.

Property *(ii)*. By local Lipschitz continuity, there is $\delta>0$ such that for all $y\in\Omega\cap B_\delta(x)$ and $\lambda\in(0,\delta)$,

$$|F(y+\lambda d)-F(y)|\leq L\lambda\|d\|.$$

The property follows by plugging this inequality into the limit in (5.5).

Property *(iii)*. This can be seen directly from the definition. □

We are now ready to define Clarke stationary points for constrained optimisation problems.

Definition 5.11 (Constrained Clarke stationary points). *Let $F:\mathbb{R}^n\to\mathbb{R}$ be locally Lipschitz continuous near $x\in\Omega$ and $d\in\mathbb{R}^n$. We say that F is* directionally Clarke optimal *at x along d if*

$$F^o(x;d)\geq 0.$$

If this holds for F at x for all $d\in T^C_\Omega(x)$, then we call x a constrained Clarke stationary point *restricted to Ω.*

This notion of first-order optimality is also considered by Audet & Dennis Jr for establishing optimality of MADS [9].

As we know from Chapter 2 in the unconstrained case, if $x\in\mathbb{R}^n$ is a local minimiser of $F:\mathbb{R}^n\to\mathbb{R}$, then $0\in\partial^C F(x)$, and for convex functions, the reverse also holds. We now present a result to show that the above notion of a constrained Clarke stationary point is consistent with these comparisons, i.e. Definition 5.11 is a sufficient optimality criteria for a wide range of optimisation problems.

For this we recall pseudoconvex functions which were defined in Chapter 2, and starshaped sets from the previous subsection. We define Clarke regular functions.

Definition 5.12 (Clarke regularity). *A function $F:\mathbb{R}^n\to\overline{\mathbb{R}}$ is* (Clarke) regular *at x if the directional derivative*

$$F'(x;d):=\lim_{\lambda\downarrow 0}\frac{F(x+\lambda d)-F(x)}{\lambda}$$

exists for all $d\in\mathbb{R}^n$ and $F^o(x;d)=F'(x;d)$. If this holds for all x, we say that F is regular.

Lemma 5.13. *Consider a function $F:\widehat{\Omega}\to\mathbb{R}$ for which $\Omega\Subset\widehat{\Omega}$, and a point $x\in\Omega$ at which F is regular, directionally differentiable, pseudoconvex, and locally Lipschitz continuous. Furthermore, suppose that Ω is starshaped with respect to x and that Ω is epi-Lipschitzian and regular at x. Then x is a Clarke stationary point of F restricted to Ω if and only if x is a global minimiser of F.*

Proof. This follows from [118, Theorem 4.19], noting that by regularity of Ω, the contingent cone and the Clarke tangent cone coincide [26, Corollary 6.1], and that by regularity of F, $F'(x;d) = F^o(x;d)$. □

5.3 Convergence of the algorithm

In Chapter 4, we considered Itoh–Abe type methods for unconstrained optimisation of nonsmooth, nonconvex functions. In Theorems 4.6, 4.10 , we proved that if the directions $(d^k)_{k\in\mathbb{N}}$ are cyclically dense or randomly drawn from S^{n-1}, then the iterates $(x^k)_{k\in\mathbb{N}}$ converge to a limit set of Clarke stationary points. In this section, we prove that under the same assumptions on the sequence of directions, the iterates of Algorithm 4 converge to a limit set of constrained Clarke stationary points of F on Ω.

First, we prove that Algorithm 4 is a well-defined and implementable scheme.

Proposition 5.14. *Let Ω be an epi-Lipschitzian subset of $\mathbb{R}^n$, and let $F : \Omega \to \mathbb{R}$ be continuous and bounded below on Ω. Then for any current iterate x^k and $d^{k+1} \in S^{n-1}$, Algorithm 4 admits a solution x^{k+1}.*

Proof. By Proposition 5.10 *(i)*, the Clarke directional derivatives $F^o(x^k;d^{k+1})$ and $F^o(x^k;-d^{k+1})$ are well-defined. If the current point is approximately directionally stationary or the constraint is (approximately) active, i.e. $F(x^k - \varepsilon_x d^{k+1}) - F(x^k) > -\varepsilon_x^2/\tau_k$, or $x^k - \varepsilon_x d^{k+1} \notin \Omega$, then set $x^{k+1} = x^k$ and we are done.

Otherwise, $x^k - \varepsilon_x d^{k+1} \in \Omega$ and $F(x^k - \varepsilon_x d^{k+1}) - F(x^k) \leq -\varepsilon_x^2/\tau_k$. In the unconstrained case, we know from Chapter 4 that a solution to (5.2) for τ_k can be found by a number of scalar root-finder methods. In the constrained case, such a method will either locate α that solves (5.2), in which case one can set $\alpha_k = \alpha$, or locate some $\alpha > 0$ such that $x^k - \alpha d^{k+1} \notin \Omega$.

It remains to consider the latter case. Set $\lambda = \alpha$ and $\alpha = \gamma\lambda$. If $x^k - \alpha d^{k+1} \in \Omega$ and $F(x^k - \alpha d^{k+1}) - F(x^k) < -\alpha^2/\tau_k$, then α solves (5.2) for $\widetilde{\tau}_k := \alpha^2/(F(x^k - \alpha d^{k+1}) - F(x^k)) \leq \tau_k$ and $\alpha \geq \gamma\lambda$, so $\alpha_k = \alpha$ is an admissible update for Algorithm 4. Else, if $x^k - \alpha d^{k+1} \notin \Omega$, then we can repeat this step, setting $\lambda = \alpha$ and $\alpha = \gamma\lambda$. After repeating this step a finite number of times, we will have $\varepsilon_x \geq \gamma\lambda$, in which case $\alpha_k = \varepsilon_x$ will be an admissible update. Finally, if $x^k - \alpha d^{k+1} \in \Omega$ and $F(x^k - \alpha d^{k+1}) - F(x^k) > -\alpha^2/\tau_k$, then by the analysis in Lemma 4.1, either there is a solution $\alpha_k \in (\varepsilon_x, \alpha)$ to (5.2) for τ_k or there is another point $\lambda \in (\varepsilon_x, \alpha)$ such that $x^k - \lambda d^{k+1} \notin \Omega$. In the former case, we are done, while in the latter case we can repeat this step a finite number of times. □

The following lemma summarises some basic properties of the method, along the lines of Lemma 4.3.

Lemma 5.15. *Consider a locally Lipschitz continuous, coercive function $F : \Omega \to \mathbb{R}$ defined on an epi-Lipschitzian set $\Omega \subset \mathbb{R}^n$, and suppose the iterates $(x^k)_{k\in\mathbb{N}}$ solve Algorithm 4. Then the following properties hold.*

(i) $F(x^{k+1}) \leq F(x^k)$.

(ii) $\|x^k - x^{k+1}\| \to 0$.

(iii) $(x^k)_{k\in\mathbb{N}}$ *has an accumulation point.*

Proof. Property (i). This follows immediately from the algorithm.

Property (ii). By coercivity and local Lipschitz continuity, F is bounded below, so we can define $F^* := \lim_{k\to\infty} F(x^k)$. We have

$$F(x^0) - F^* = \sum_{k\in\mathbb{N}} F(x^k) - F(x^{k+1}) = \sum_{k\in\mathbb{N}} \frac{1}{\tau_k}\|x^k - x^{k+1}\| \geq \frac{1}{\tau_{\max}} \sum_{k\in\mathbb{N}} \|x^k - x^{k+1}\|.$$

Therefore $\|x^k - x^{k+1}\| \to 0$.

Property (iii). This follows from coercivity of F. □

We denote by the limit set S of $(x^k)_{k\in\mathbb{N}}$ as given in (4.2), which is nonempty by the above lemma. The following result is a standard extension of Lemma 4.4 and the same proof holds.

Lemma 5.16. *The limit set S is compact, connected and has empty interior. Furthermore, F is constant on S.*

5.3.1 Stochastic case

We proceed to the convergence proofs. We first prove convergence when $(d^k)_{k\in\mathbb{N}}$ are chosen stochastically. As in the previous chapter, we suppose that $(d^k)_{k\in\mathbb{N}}$ are randomly, independently drawn, and that the support of the probability density of Ξ is dense in S^{n-1}.

We will make use of the following results from [9], which extend the Lipschitz continuity of the Clarke generalised derivative to the constrained case for the hypertangent cone, and continuity to its closure, the Clarke tangent cone.

Proposition 5.17 ([9, Lemma 3.8])**.** *Let F be Lipschitz continuous with Lipschitz constant $L > 0$ near $x \in \Omega$. If d and e are in $T_\Omega^H(x)$, then*

$$|F^o(x;d) - F^o(x;e)| \leq L\|d - e\|.$$

Proposition 5.18 ([9, Proposition 3.9])**.** *Let F be Lipschitz continuous near $x \in \Omega$. If Ω is epi-Lipschitzian at x and $d \in T_\Omega^{\mathrm{Cl}}(x)$, then*

$$F^o(x;d) = \lim_{\substack{e \to d, \\ e \in T_\Omega^H(x)}} F^o(x;e).$$

Note that it follows from Proposition 5.18 that if Ω is epi-Lipschitzian at x, then for Clarke stationarity on Ω, it is sufficient to verify optimality on $T_\Omega^H(x)$ rather than $T_\Omega^C(x)$.

We define X to be the set of nonstationary points,

$$X = \{x \in \Omega \,:\, F \text{ is not Clarke stationary at } x \text{ restricted to } \Omega\}. \tag{5.6}$$

Theorem 5.19. *Let $F : \Omega \to \mathbb{R}$ be locally Lipschitz continuous, coercive, and bounded below, where $\Omega \subset \mathbb{R}^n$ is epi-Lipschitzian. Let $(x^k)_{k\in\mathbb{N}}$ solve Algorithm 4 for $\varepsilon_x = \varepsilon_F = 0$, where $(d^k)_{k\in\mathbb{N}}$ are independently drawn from the random distribution Ξ, and suppose that the support of the density of Ξ is dense in S^{n-1}. Then $\mathbb{P}(S \cap X \neq \varnothing) = 0$, i.e. the limit set S is almost surely in the set of stationary points.*

Proof. This proof is analogous to the proof for the unconstrained case, with the exception of additional treatment of points satisfying the constraints, and the progress parameter γ.

We will construct a countable collection of open sets $(B_j)_{j\in\mathbb{N}}$, such that $X \subset \bigcup_{j\in\mathbb{N}} B_j$ and so that for all $j \in \mathbb{N}$ we have $\mathbb{P}(S \cap B_j \neq \varnothing) = 0$. Then the result follows from countable additivity of probability measures.

First, we show that for every $x \in X$, there is $d \in S^{n-1} \cap T_\Omega^H(x)$, $\varepsilon > 0$, and $\delta > 0$ such that

$$y - \lambda e \in \Omega,\ \frac{F(y-\lambda e) - F(y)}{\lambda} \leq -\varepsilon, \quad \forall y \in B_\delta(x),\ e \in B_\delta(d) \cap S^{n-1},\ \lambda \in (0,\delta). \tag{5.7}$$

To show this, note that if $x \in X \subset \Omega$, then as Ω is epi-Lipschitzian, there is $d \in S^{n-1} \cap T_\Omega^H(x)$ and $\varepsilon > 0$ such that

$$F^o(x;-d) = \limsup_{\substack{y\to x,\ \lambda \downarrow 0 \\ y \in \Omega,\ y+\lambda d \in \Omega}} \frac{F(y-\lambda d) - F(y)}{\lambda} \leq -\varepsilon.$$

Therefore, there is $\eta > 0$ such that for all $\lambda \in (0,\eta)$ and all $y \in B_\eta(x)$, we have

$$\frac{F(y-\lambda d) - F(y)}{\lambda} \leq -\varepsilon/2.$$

As F is Lipschitz continuous around $B_\eta(x)$, it is clear that the mapping

$$e \mapsto \frac{F(y-\lambda e)-F(y)}{\lambda},$$

is also locally Lipschitz continuous. By this and since $d \in T_\Omega^H(x)$, it follows that there exists $\delta \in (0,\eta)$ such that for all $y \in B_\delta(x) \cap \Omega$, all $e \in B_\delta(d) \cap S^{n-1}$, and all $\lambda \in (0,\delta)$, we have

$$y-\lambda e \in \Omega, \text{ and } \frac{F(y-\lambda e)-F(y)}{\lambda} \leq -\varepsilon/3.$$

This concludes the first part.

Next, for $m \in \mathbb{N}$, we define the set

$$X_m = \left\{ x \in X \;:\; (5.7) \text{ holds for some } d \in S^{n-1} \cap T_\Omega^H(x),\ \varepsilon > 0, \text{ and all } \delta < 1/m \right\}.$$

Clearly

$$X = \bigcup_{m \in \mathbb{N}} X_m.$$

Let $(y^i)_{i \in \mathbb{N}}$ be a dense sequence in X_m, which exists because $\mathbb{Q}^n$ is both countable and dense in Ω. We define $Y_i^{(m)} = B_\delta(y^i)$, where $\delta = \frac{1}{m+1}$. Therefore,

$$X_m \subset \bigcup_{i \in \mathbb{N}} Y_i^{(m)} \quad \Longrightarrow \quad X \subset \bigcup_{m \in \mathbb{N}} \bigcup_{i \in \mathbb{N}} Y_i^{(m)}.$$

Since a countable union of countable sets is countable, we conclude with the following statement. For each $i \in \mathbb{N}$ there is $y^i \in \Omega$, $\varepsilon_i > 0$, $\delta_i > 0$, and $\widetilde{d^i} \in S^{n-1} \cap T_\Omega^H(y^i)$, such that for all $z \in B_{\delta_i}(y^i)$, all $\widetilde{d} \in B_{\delta_i}(\widetilde{d^i}) \cap S^{n-1}$, and all $\lambda \in (0,\delta_i)$, we have

$$z-\lambda\widetilde{d} \in \Omega, \quad \frac{F(z-\lambda\widetilde{d})-F(z)}{\lambda} \leq -\varepsilon_i,$$

and such that

$$X \subset \bigcup_{i \in \mathbb{N}} B_{\delta_i}(y^i). \tag{5.8}$$

Finally, we show that for each $i \in \mathbb{N}$, almost surely, $S \cap B_{\delta_i}(y^i) = \varnothing$. For a given i, write $B_i := B_{\delta_i}(y^i)$, and define $m := \min_{x \in B_i} F(x)$, $M := \max_{x \in B_i} F(x)$. We argue accordingly: The existence of an accumulation point of $(x^k)_{k \in \mathbb{N}}$ in B_i would imply that there is a subsequence $(x^{k_j})_{j \in \mathbb{N}} \subset B_i$. Suppose $x^{k_j} \in B_i$ and $d^{k_j+1} \in B_{\delta_i}(\widetilde{d^i})$. Then $x^{k_j+1} = x^{k_j} - \alpha_{k_j} d^{k_j+1}$, where

α_{k_j} either solves (5.2) for τ_{k_j} or solves (5.2) for some $\widetilde{\tau}_{k_j} \in (0, \tau_{k_j}]$ and such that $\alpha_{k_j} \geq \gamma\delta_i$ (since any λ such that $x^k - \lambda d^{k+1} \notin \Omega$ must be greater than δ_i).

If α_{k_j} solves (5.2) for τ_{k_j}, then by the analysis in the proof of Theorem 4.6, we know that

$$F(x^{k_j}) - F(x^{k_j+1}) \geq \min\left\{\varepsilon_i^2\tau_{\min}, \frac{\delta_i^2}{\tau_{\max}}\right\}.$$

Otherwise, $\alpha_{k_j} > \gamma\delta_i$ and there is $\widetilde{\tau}_{k_j} \in (0, \tau_{\max}]$ such that $x^{k_j+1} = x^{k_j} + \alpha_{k_j} d^{k_j+1}$ and

$$F(x^{k_j}) - F(x^{k_j+1}) = \frac{1}{\widetilde{\tau}_{k_j}}\alpha_{k_j}^2.$$

In this case,

$$F(x^{k_j}) - F(x^{k_j+1}) = \frac{1}{\widetilde{\tau}_{k_j}}\alpha_{k_j}^2 \geq \frac{1}{\tau_{\max}}\alpha_{k_j}^2 \geq \frac{1}{\tau_{\max}}\gamma^2\delta_i^2.$$

Setting $\mu = \min\{\varepsilon_i^2\tau_{\min}, \delta_i^2/\tau_{\max}, \gamma^2\delta_i^2/\tau_{\max}\}$, it follows that whenever $x^{k_j} \in B_i$ and $d^{k_j+1} \in B_{\delta_i}(\widetilde{d^i})$, then

$$F(x^{k_j}) - F(x^{k_j+1}) \geq \mu.$$

Choosing $K \in \mathbb{N}$ such that $K\mu > M - m$, we know that this event only has to occur K times for $(x^{k_j})_{j\in\mathbb{N}}$ to leave B_i. In other words, almost surely, there is no subsequence $(x^{k_j})_{j\in\mathbb{N}} \subset B_i$. This concludes the proof. □

Deterministic case

We now state the deterministic case, in which $(d^k)_{k\in\mathbb{N}}$ is required to be cyclically dense. Its proof is simply that of Theorem 4.10, but referring to the proof of Theorem 5.19 instead of Theorem 4.6.

Theorem 5.20. *Let $F : \Omega \to \mathbb{R}$ be locally Lipschitz continuous, coercive, and bounded below, where $\Omega \subset \mathbb{R}^n$ is epi-Lipschitzian. Let $(x^k)_{k\in\mathbb{N}}$ solve Algorithm 4 for $\varepsilon_x = \varepsilon_F = 0$, where $(d^k)_{k\in\mathbb{N}}$ are cyclically dense. Then the limit set S is in the set of stationary points.*

5.4 Numerical experiments

We consider some simple numerical examples on $\mathbb{R}^2$. Algorithm 4 has been implemented on MATLAB, using a simple bisectional search method to solve the scalar equation. For the algorithmic parameters, we have chosen $\tau_k = 1$ for all $k \in \mathbb{N}$, d^k drawn independently from

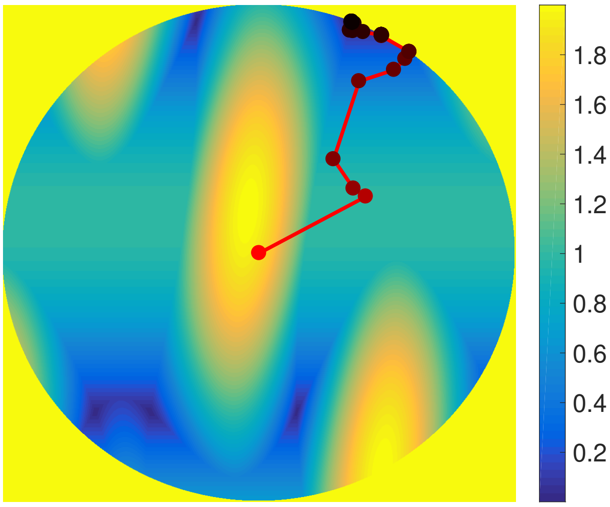


Fig. 5.1 Numerical results for the optimisation problem in (5.9), with iterates going from red to black.

the uniform distribution on S^{n-1}, $\gamma = 0.5$, $\varepsilon_x = 10^{-5}$, and $\varepsilon_F = 10^{-5}$. The algorithm is set to stop if the objective value has decreased by less than ε_F over the last 30 iterates.

We first consider the optimisation problem (5.1) with $F : \mathbb{R}^2 \to \mathbb{R}$ and Ω given by

$$\begin{aligned} F(x) &:= \max\{|\cos(x_1+x_2)+\sin(3x_2)|, |\sin(x_1+1)|\}, \\ \Omega &:= \{x \in \mathbb{R}^2 \ : \ (x_1-4)^2+(x_2-2.7)^2 \leq 4\}. \end{aligned} \tag{5.9}$$

The results are plotted in Figure 5.1, with increasing iterates plotted with darker colours, and the infeasible region plotted in yellow.

Next, we consider the optimisation problem with nonsmooth, nonconvex constraints, given by

$$\begin{aligned} F(x) &:= \max\{|\cos(x_1+x_2+3)+\sin(3x_2-0.5)|, |\sin(x_1+4.5)|\}, \\ \Omega &:= \{-2x_1+1.2x_2 \leq -4\} \cup \{x_1+x_2 \geq 6\} \cup \{x_1+2x_2 \leq 11.5\} \ldots \\ &\qquad \cap \{-2x_1+x_2 \geq -7\} \cap \{-x_1+x_2 \leq -1\} \end{aligned} \tag{5.10}$$

The results are plotted in Figure 5.2.

5.5 Conclusion and outlook

In this chapter, we have extended the Itoh–Abe methods to constrained optimisation problems, and proven convergence guarantees to Clarke stationary points in this setting. This extension and analysis is important because bilevel problems and simulation-based parameter

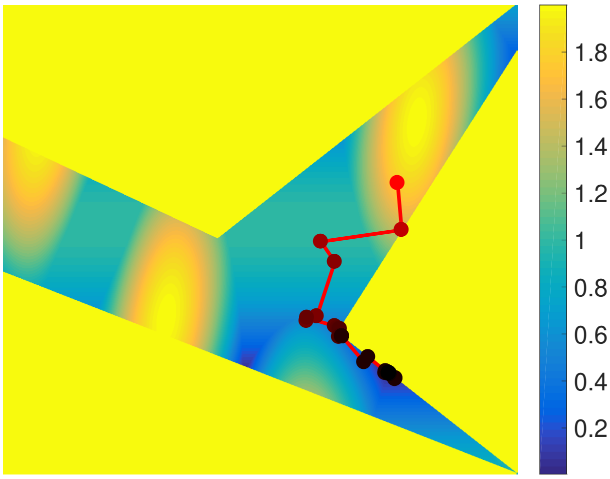


Fig. 5.2 Numerical results for the optimisation problem in (5.10), with iterates going from red to black.

optimisation problems, which constitute a central motivation for the study of black-box optimisation methods, often feature constraints in the domain. Furthermore, these constraints might be implicitly defined so that we have no information of the feasible set beyond verifying feasibility point by point. It is therefore important to treat this in the nonsmooth, nonconvex, derivative-free optimisation framework. We apply these methods to solve some simple, numerical examples. A wider numerical investigation of this approach is left for future work.

Chapter 6

Bregman discrete gradient methods for sparse optimisation

6.1 Introduction

This chapter is based on the article [20] published in the Journal of Mathematical Imaging and Vision, and is joint work with Martin Benning and Carola-Bibiane Schönlieb.

In Chapters 3–5, we studied the discrete gradient method applied to gradient flow in various optimisation settings. In this chapter, we study these methods applied to a different dissipative flow, namely the *inverse scale space* (ISS) flow.

We consider the constrained optimisation problem

$$\min_{x\in\Omega} F(x), \tag{6.1}$$

for an objective function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and constraint $\Omega \subset \mathbb{R}^n$. The function F may be nonconvex and nonsmooth, as outlined in Assumption 6.1. In this chapter, we propose and study discrete gradient methods applied to the ISS flow.

The ISS flow is a differential system which generalises gradient flows by replacing the Euclidean distance by a Bregman distance, defined via a convex Bregman distance generating function $J : \mathbb{R}^n \to \overline{\mathbb{R}}$. The ISS flow is given by

$$\dot{p}(t) = -\nabla F(x(t)), \quad p(t) \in \partial J(x(t)). \tag{6.2}$$

The term *inverse scale space flow* goes back to Scherzer & Groetsch [201]. It is typically derived as the continous-time limit of Bregman iterations (2.3). Like the gradient flow, the ISS flow is a dissipative system, and its dissipative structure is determined by the function J.

This allows one to solve (6.1) while incorporating *a priori* information into the optimisation scheme, with the benefits of converging to superior solutions, and doing so faster. The drawback of these methods is that the updates are in general implicit. Nevertheless, for many simple variational problems, the updates turn out to be explicit.

In this chapter, we study the Itoh–Abe discrete gradient method applied to the ISS flow. We prove that the method is well-defined and converges to a set of stationary points for nonsmooth, nonconvex functions. Furthermore, building on the paper by Miyatake et al. [153] where they establish the equivalence between the discrete gradient methods for linear systems and successive-over-relaxation (SOR) methods, we point out equivalencies of various approaches to least squares problems.

Bregman iterations, and related methods, are closely tied to inverse problems and regularisation methods, particularly in signal processing. We consider numerical examples in this setting as well.

6.1.1 Related literature

Spurred by applications for variational regularisation in image processing and compressed sensing, the ISS flow and the Bregman method have been active areas of research during the last decade. The Bregman iterative method was originally proposed by Osher et al. [170] in 2005 for total variation-based image denoising, representing an extension of the Bregman proximal algorithm [46, 78, 122, 211] to nonsmooth Bregman distance generating functions. Subsequently the ISS flow was derived and analysed by Burger et al. [34, 37, 35, 32]. Since then, researchers have studied the ISS flow with applications to generalised spectral analysis in a nonlinear setting, i.e. by Burger et al. [33], Gilboa et al. [94], and Schmidt et al. [202]. The Bregman method has been studied for ℓ^1-regularisation and compressed sensing by Goldstein & Osher [96] and Yin et al. [223], and extended to primal-dual algorithms by Zhang et al. [225].

The linearised Bregman method was proposed by Yin et al. [223] for applications to ℓ^1-regularisation and compressed sensing, and further studied in this setting by Cai et al. [39], and Dong et al. [72]. An extension for nonconvex problems was proposed by Benning et al. [17], proving global convergence for functions that satisfy the Kurdyka–Łojasiewicz property. Lorenz et al. [142, 203] proposed a sparse variant of the Kaczmarz method for linear problems based on linearised Bregman iterations. These and other methods were unified in a Split Feasibility Problems framework for general convergence results by Lorenz et al. [141]. For further details on Bregman iterations and linearised Bregman methods, we refer to [18].

6.1.2 Structure and contributions

The rest of the chapter is structured as follows. In Section 6.2, we introduce the ISS flow and propose a Bregman discrete gradient method based on the ISS flow. In Section 6.3 we prove well-posedness and convergence results in a nonconvex, nonsmooth framework. In Section 6.4, we discuss particular examples of Bregman discrete gradient methods, while in Section 6.5 we establish equivalencies to other optimisation schemes. In Section 6.6, we present results from numerical experiments.

6.2 The discrete gradient method for the ISS flow

6.2.1 Inverse scale space flow and Bregman methods

For a convex function $J : \mathbb{R}^n \to \overline{\mathbb{R}}$, objective function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ and starting points $x(0) = x^0 \in \Omega$, $p(0) \in \partial J(x^0)$, the ISS flow is the dissipative differential system given by (6.2). If J were twice continuously differentiable and μ-convex, then (6.2) could be rewritten as

$$\dot{x}(t) = -(\nabla^2 J(x(t)))^{-1} \nabla F(x(t)),$$

and the energy $F(x(t))$ would dissipate over time as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} F(x(t)) &= \langle \dot{x}(t), \nabla F(x(t)) \rangle \\ &= -\left\langle \dot{x}(t), \nabla^2 J(x(t)) \dot{x}(t) \right\rangle \leq -\mu \|\dot{x}(t)\|^2.\end{aligned}$$

We briefly discuss variants of Bregman methods as discretisations of (6.2). The Bregman method is derived by backward Euler discretisation of (6.2), and is given by

$$p^{k+1} = p^k - \tau_k \nabla F(x^{k+1}), \quad p^{k+1} \in \partial J(x^{k+1})$$

which can be rewritten as

$$x^{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} F(x) + \frac{1}{\tau_k} D_J^{p^k}(x, x^k). \tag{6.3}$$

From (6.3), we see that the Bregman method is dissipative, as

$$F(x^{k+1}) - F(x^k) \leq -\frac{1}{\tau_k} D_J^{p^k}(x^{k+1}, x^k) \leq -\frac{\mu}{2\tau_k} \|x^k - x^{k+1}\|^2.$$

Similarly, the linearised Bregman method is derived by forward Euler discretisation of (6.2), and is given by

$$p^{k+1} = p^k - \tau_k \nabla F(x^k), \quad p^{k+1} \in \partial J(x^{k+1})$$

or equivalently

$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\arg\min} F(x^k) + \langle \nabla F(x^k), x - x^k \rangle + \frac{1}{\tau_k} D_J^{p^k}(x, x^k).$$

The ISS flow and Bregman methods are considered for solving ill-conditioned linear systems $Ax = b$, with the objective function

$$F(x) = \frac{1}{2}\|Ax - b\|^2.$$

In this setting, iterates of both the Bregman method and the linearised Bregman method converge [18, 141] to a solution of

$$\min_{x \in \mathbb{R}^n} \{J(x) \text{ s.t. } Ax = b\}.$$

Furthermore, applications of the ISS flow include image denoising with reduced contrast-loss and staircasing effects [170], recovering eigenfunctions [202], and identifying sparse or low-rank structures [223].

We make the following assumptions for the objective function F, the constraints Ω, and the Bregman distance generating function J.

Assumption 6.1.

a) The function $F : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous and bounded below.

b) $x^* \in \mathbb{R}^n$ is a Clarke stationary point of F restricted to Ω if and only if for all coordinate vectors e^i, we have $F^o(x^*; e^i), F^o(x^*; -e^i) \geq 0$.

c) The set $\Omega \subset \mathbb{R}^n$ consists of coordinate-wise box constraints, i.e. $\Omega = \otimes_{i=1}^n [l_i, u_i]$.

d) The function $J : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lower-semicontinuous, and μ-convex with $\mu > 0$. Furthermore, $J(x) = \sum_{i=1}^n j_i(x_i) + \delta_{[l_i,u_i]}(x_i)$, where $[l_i, u_i] \subset \mathrm{dom}(j_i)$ for each i.

6.2.2 The Bregman discrete gradient method

In what follows, we define discrete gradients, and propose the Bregman discrete gradient method by discretising the ISS flow. For $i = 1, \dots, n$, we denote by $[\partial F(x)]_i$ the projection of $\partial F(x)$ onto the ith coordinate, i.e. $[\partial F(x)]_i = \{p_i \, : \, p \in \partial F(x)\}$.

We propose the *Bregman discrete gradient method* as follows. For a starting point $x^0 \in \mathbb{R}^n$, subgradient $p^0 \in \partial J(x^0)$, and time steps $(\tau_{k,i})_{i=1}^n \subset [\tau_{\min}, \tau_{\max}]^n$, solve for $k = 0, 1, \dots$,

$$p^{k+1} = p^k - \tau_k \overline{\nabla} F(x^k, x^{k+1}), \qquad p^{k+1} \in \partial J(x^{k+1}). \tag{6.4}$$

This scheme preserves the dissipative structure of the ISS flow and (linearised) Bregman methods, as we see by applying the mean value property (2.12) and (6.4).

$$\begin{aligned} F(x^k) - F(x^{k+1}) &= \langle x^k - x^{k+1}, \overline{\nabla} F(x^k, x^{k+1}) \rangle \\ &= \frac{1}{\tau_k} \langle x^k - x^{k+1}, p^k - p^{k+1} \rangle \\ &= \frac{1}{\tau_k} D_J^{\text{symm}}(x^{k+1}, x^k) \\ &\geq \frac{\mu}{\tau_k} \|x^k - x^{k+1}\|^2. \end{aligned} \tag{6.5}$$

Furthermore, if we plug in $J(x) = \|x\|^2/2$, we recover the discrete gradient method for the gradient flow (2.8).

By Assumption 6.1, the subdifferential of J is separable in the coordinates, i.e.

$$\partial J(x) = \prod_{i=1}^n \partial \delta_{[l_i, u_i]}(x_i) + \partial j_i(x_i).$$

It follows that solving the Bregman Itoh–Abe equation (6.4) corresponds to successively solving n scalar equations,

$$\begin{aligned} p_i^{k+1} + q_i^{k+1} &= p_i^k - \tau_{k,i} \frac{F(y^{k,i}) - F(y^{k,i-1})}{x_i^{k+1} - x_i^k}, \\ p_i^{k+1} \in \partial j_i(y_i^{k,i}), &\quad q_i^{k+1} \in \partial \delta_{[l_i, u_i]}(y_i^{k,i}), \\ y^{k,i} &= [x_1^{k+1}, \dots, x_i^{k+1}, x_{i+1}^k, \dots, x_n^k], \quad i = 1, \dots, n. \end{aligned} \tag{6.6}$$

Here $y^{k,i}$ denotes $[x_1^{k+1}, \dots, x_i^{k+1}, x_{i+1}^k, \dots, x_n^k]^T$. For a choice of $v_i^k \in [\partial F(y^{k,i-1})]_i$, if $p_i^k - \tau_{k,i} v_i^k \in [\partial J(y^{k,i-1})]_i$, then we consider $x_i^{k+1} = x_i^k$ and $p_i^{k+1} + q_i^{k+1} = p_i^k - \tau_{k,i} v_i^k$ an admissible update.

We include the term q^{k+1} to absorb subdifferential updates due to active constraints, i.e. $\partial \delta_\Omega$, and to not include them in the next update. The purpose of this is first to prevent the dual variables to diverge while the primal variables are unchanged, and second to obtain guarantees of first-order optimality for accumulation points of the iterates. Additionally, since $0 \in \partial \delta_\Omega(x)$ for all $x \in \Omega$, we still have $p^{k+1} \in \partial J(x^{k+1})$ and (6.5) still holds.

6.3 Well-posedness and convergence

In this section, we prove that the Bregman discrete gradient method (6.6) is well-defined. Furthermore, we prove that all accumulation points of the iterates $(x^k)_{k\in\mathbb{N}}$ defined via (6.6) are Clarke stationary points.

6.3.1 Well-posedness

Lemma 6.2. *For any $\tau > 0$, $x^k \in \mathbb{R}^n$, and $p^k \in \partial J(x^k)$, there exists an update $(x^{k+1}, \tilde{p}^{k+1} = p^{k+1} + q^{k+1})$ that satisfies* (6.6)*.*

Proof. As (6.6) consists of successive scalar updates, it is sufficient to consider a scalar problem, $v : \mathbb{R} \to \mathbb{R}$, $j : \mathbb{R} \to \overline{\mathbb{R}}$. For $x \in \mathbb{R}$ and $p \in \partial j(x)$ we either want $y \neq x$ such that

$$p - \tau \frac{v(y) - v(x)}{y - x} \in \partial j(y), \tag{6.7}$$

or $y = x$ and $p - \tau w \in \partial j(x)$, for some $w \in \partial v(x)$.

If such a w exists, we are done. Otherwise, we have $\min\{v^o(x;1), v^o(x;-1)\} < 0$ and may assume that $v^o(x;1) < 0$. In this case, we will show that there exists $y > x$ such that (6.7) holds.

Since $p - \tau v^o(x;1) > p$ and $p \in \partial j(x)$, we deduce that $p - \tau v^o(x;1) > \partial j(x)$. By the outer semicontinuity of subdifferentials and definition of Clarke directional derivatives, there is $\delta > 0$ such that

$$p - \tau \frac{v(y) - v(x)}{y - x} > \partial j(y) \text{ for all } y \in (x, x + \delta).$$

On the other hand, as v is bounded below,

$$y \mapsto (v(y) - v(x))/(y - x)$$

is bounded below for all $y \in [x+\delta,+\infty)$, while by μ-convexity of j, we have $\partial j(y) \geq \partial j(x) + \mu(y-x)$ for all $y \in [x+\delta,+\infty)$. Hence, there is $r \gg 0$ such that

$$p - \tau \frac{v(y)-v(x)}{y-x} < \partial j(y) \text{ for all } y \geq x+r.$$

By continuity of v, and by outer semicontinuity of subdifferentials, it follows that there exists $y \in (x+\delta, x+r)$ that solves (6.7). This concludes the proof. □

Corollary 6.3. *If $F : \mathbb{R}^n \to \mathbb{R}$ is convex, then there exists a unique solution $(x^{k+1}, \tilde{p}^{k+1} = p^{k+1} + q^{k+1})$ to* (6.6).

Proof. The existence of a solution to (6.6) is guaranteed by Lemma 6.2. To establish uniqueness, we argue as follows. An update $y^{k,i}$ must satisfy

$$p_i^k - \tau_{k,i} \frac{F(y^{k,i}) - F(y^{k,i-1})}{x_i^{k+1} - x_i^k} \in [\partial J(y^{k,i})]_i.$$

The left-hand-side is non-increasing with respect to x_i^{k+1}, due to the difference quotient term of a convex function F, while the right-hand side is strictly increasing, due to the strong convexity of J. Hence there cannot be two distinct solutions for $y_i^{k,i}$ to the scalar equation. This implies uniqueness of the update. □

Remark 6.4. *If the update is stationary, i.e. $x_i^{k+1} = x_i^k$, then the subgradient update p_i^{k+1} is unique only up to the choice of subderivative $v_i \in [\partial F(y^{k,i-1})]_i$.*

6.3.2 Convergence theorem

Lemma 6.5. *Let $F : \mathbb{R}^n \to \mathbb{R}$, $J : \mathbb{R}^n \to \overline{\mathbb{R}}$, and Ω satisfy Assumption 6.1, and let $(x^k, p^k)_{k\in\mathbb{N}}$ be iterates that solve* (6.6) *for time steps $(\tau_k)_{k\in\mathbb{N}} \subset [\tau_{\min}, \tau_{\max}]$. Then the following properties hold.*

(i) $F(x^{k+1}) \leq F(x^k)$.

(ii) $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$.

(iii) If F is coercive, then there exists a convergent subsequence of $(x^k, p^k)_{k\in\mathbb{N}}$.

(iv) The set of limit points S is compact, connected, and has empty interior. Furthermore, F is single-valued on S.

Proof. Property *(i)* follows from (6.5).

As F is bounded below and $(F(x^k))_{k\in\mathbb{N}}$ is decreasing, $F(x^k)\to F^*$ for some limit F^*. Therefore, by (6.5),

$$\begin{aligned}F(x^0)-F^* &= \sum_{k=0}^{\infty} F(x^k)-F(x^{k+1})\\ &\geq \sum_{k=0}^{\infty}\frac{\mu}{\tau_k}\|x^k-x^{k+1}\|^2 \geq \frac{\mu}{\tau_{\max}}\sum_{k=0}^{\infty}\|x^k-x^{k+1}\|^2.\end{aligned}$$

This implies property *(ii)*.

Properties *(iii)* and *(iv)* follow from *(i)* and *(ii)* and are proven respectively in Lemma 4.3 and Lemma 4.4. □

We now state and prove the main result of this chapter.

Theorem 6.6. *Let the sequence of iterates* $(x^k,p^k)_{k\in\mathbb{N}}$ *solve* (6.6) *for time steps* $(\tau_k)_{k\in\mathbb{N}}\subset[\tau_{\min},\tau_{\max}]$. *Then all accumulation points* $x^*\in S$ *are Clarke stationary points restricted to* Ω.

Proof. Let $x^*\in S$ and consider a convergent subsequence $(x^{k_j})_{j\in\mathbb{N}}$. We want to show for each basis vector e^i that either $F^o(x^*;e^i)\geq 0$ or $x_i^*=u_i$, and analogously that either $F^o(x^*;-e^i)\geq 0$ or $x_i^*=l_i$. As the arguments are identical, we only consider the first case.

Suppose for contradiction that $F^o(x^*;e^i)<-\eta$ for some $\eta>0$, and that $x_i^*<u_i$. By the definition of the Clarke directional derivative, there are $\varepsilon,\delta>0$ such that for all $x\in B_\varepsilon(x^*)$ and $\lambda\in(0,\delta)$, we have

$$\frac{F(x+\lambda e^i)-F(x)}{\lambda}\leq-\frac{\eta}{2}. \tag{6.8}$$

Since $x^{k_j}\to x^*$ and $\|x^{k_j+1}-x^{k_j}\|\to 0$, for each $N\in\mathbb{N}$ there exists K such that for all $j\geq K$, we have $x^k\in B_\varepsilon(x^*)$ and $\|x^k-x^{k+1}\|<\delta$ for $k=k_j,k_j+1,\dots,k_j+N$. By making $\varepsilon>0$ sufficiently small, we have $B_\varepsilon(x_i^*)<u_i$. Furthermore, since $x_i^{k+1}\geq x_i^k$ for $k=k_j,\dots,k_j+N-1$, we deduce that the constraint component q_i^k is zero. By (6.8), it follows that

$$\begin{aligned}p_i^{k_j}-p_i^{k_j+N} &= \sum_{k=k_j}^{N-1}p_i^k-p_i^{k+1}=\sum_{k=k_j}^{N-1}\tau_i^k\frac{F(y^{k,i})-F(y^{k,i-1}))}{x_i^{k+1}-x_i^k}\\ &\leq-\tau_{\min}\sum_{k=k_j}^{N-1}\frac{\eta}{2}=-N\tau_{\min}\frac{\eta}{2}.\end{aligned} \tag{6.9}$$

By Assumption 6.1, ∂j_i is bounded on $U=B_\varepsilon(x^*)\cap[l_i,u_i]$. Since $p_i^{k,j},\dots,p_i^{k_j+N}\in\partial j_i(U)$, we can choose N such that $N\tau_{\min}\frac{\eta}{2}>\max\partial j_i(U)-\min\partial j_i(U)$ and arrive at a contradiction. Thus, x^* is a Clarke stationary point restricted to Ω. □

6.4 Examples of Bregman discrete gradient schemes

In this section, we describe several schemes based on the Bregman Itoh–Abe discrete gradient scheme (6.6). We will primarily consider objective functions of the form

$$F(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle, \tag{6.10}$$

where $A \in \mathbb{R}^{n\times n}$ is a positive semi-definite matrix.

We are particularly interested in problems with underlying sparsity and/or constraints, with applications in image analysis. Throughout this section, we use a time step vector τ^k coordinate-wise scaled by the diagonal of A, i.e. $\tau^k = \tau/\operatorname{diag}(A) = [\tau/a_1^1, \ldots, \tau/a_n^n]$ for all $k \in \mathbb{N}$, and some $\tau > 0$.

We first introduce some well-known coordinate descent schemes for solving linear systems, which Miyatake et al. [153] showed were equivalent to the Itoh–Abe discrete gradient method. The SOR method [224] updates each coordinate sequentially according to the rule

$$\begin{aligned} y^{k,0} &= x^k \\ y^{k,i} &= y^{k,i-1} - \frac{\omega}{a_i^i}(\langle a^i, y^{k,i-1}\rangle - b_i)e^i, \\ x^{k+1} &= y^{k,n}, \end{aligned} \tag{6.11}$$

where $\omega \in (0,2)$. For $\omega = 1$, this is the Gauss-Seidel method [224]. The SOR method is equivalent to the Itoh–Abe discrete gradient method

$$x^{k+1} = x^k - \tau\overline{\nabla}F(x^k, x^{k+1}),$$

with F given by (6.10) with the time steps $\tau_i = 2\omega/\left((2-\omega)a_i^i\right)$.

6.4.1 Sparse SOR method

We consider underdetermined linear systems and want to find sparse solutions x^*. Hence we seek to apply the Bregman discrete gradient method (6.6) with objective function F given by (6.10), and

$$J(x) = \frac{1}{2}\|x\|^2 + \gamma\|x\|_1, \tag{6.12}$$

for $\gamma > 0$. We term this the *Bregman SOR (BSOR) method.*

By Corollary 6.3, the updates of this method are well-defined and unique. One can verify that the updates are given as follows. Denote by $\tilde{x}_i^{k+1}$ the standard SOR coordinate update from x_i^k, (6.11). Furthermore, for $p^k \in \partial J(x^k)$, we write $p^k = x^k + \gamma r^k$, where $r^k \in \partial\|x^k\|_1$.

Then (x_i^{k+1}, r_i^{k+1}) are given in closed form as

$$\begin{aligned} x_i^{k+1} &= S\left(\tilde{x}_i^{k+1} + \frac{2\gamma}{2+\tau} r_i^k, \frac{2\gamma}{2+\tau}\right), \\ r_i^{k+1} &= r_i^k + \frac{\tau}{\gamma a_i^i}\left(b_i - \langle a^i, x^k\rangle - \frac{a_i^i(2+\tau)}{2\tau}(y_i - x_i)\right), \end{aligned} \tag{6.13}$$

where S is the shrinkage operator (1.5).

6.4.2 Sparse, regularised SOR

If $b = Ax^{\text{true}} + \delta$, where x^{true} is the sparse ground truth and δ is noise, then it may be necessary to regularise the objective function as well. Hence we consider the objective function

$$F(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle + \lambda \|x\|_1, \tag{6.14}$$

for some regularisation parameter $\lambda > 0$. The nonsmoothness induced by $\|\cdot\|_1$ satisfies Assumption 6.1, so Theorem 6.6 implies that the Bregman Itoh–Abe discrete gradient method converges to stationary points of this problem.

For both $J(x) = \frac{1}{2}\|x\|^2 + \gamma\|x\|_1$ and $J(x) = \frac{1}{2}\|x\|^2$, the scheme (6.6) can be expressed in closed form for (6.14), albeit with a lengthy case-by-case analysis. Consider the Bregman Itoh–Abe method with F given in (6.14) and J given in (6.12), and denote by $\tilde{x}_i^{k+1}$ the standard SOR update (6.11) for the ith coordinate. Then the ℓ^1-regularised sparse SOR method can be expressed as follows.

1. If $x_i^k = 0$ and $|\tilde{x}_i^{k+1} - \gamma r_i^k| \leq \gamma + \lambda\tau/a_i^i$, then

$$x_i^{k+1} = 0, \quad r_i^{k+1} = \frac{\gamma r_i^k - \tilde{x}_i^{k+1}}{\gamma + \lambda\tau/a_i^i}.$$

2. Else if

$$|(\tau/2+1)\tilde{x}_i^{k+1} + \gamma r_i^k| \geq \gamma + \tau\lambda/a_i^i,$$

then

$$\begin{aligned} x_i^{k+1} &= \tilde{x}_i^{k+1} + \frac{\gamma r_i^k - \left(\gamma + \tau\lambda/a_i^i\right)\operatorname{sgn}\left(\tilde{x}_i^{k+1} + \frac{\gamma}{\tau/2+1} r_i^k\right)}{\tau/2+1} \\ r_i^{k+1} &= \operatorname{sgn}\left(\tilde{x}_i^{k+1} + \frac{\gamma r_i^k}{(\tau/2+1)}\right). \end{aligned}$$

3. Else if $x_i^k \neq 0$ and

$$\Big|\big(\tau/2+1\big)\,\tilde{x}_i^{k+1}+\gamma r_i^k-(\lambda\,\tau/a_i^i)\,\mathrm{sgn}(x_i^k)\Big|\leq\gamma,$$

then set

$$\begin{aligned} x_i^{k+1} &= 0,\\ r_i^{k+1} &= r_i^k+\frac{1}{\gamma}\Big(\big(\tau/2+1\big)\,\tilde{x}_i^{k+1}-(\tau\lambda/a_i^i)\,\mathrm{sgn}(x_i^k)\Big). \end{aligned}$$

4. Else if $x_i \neq 0$ and

$$\begin{aligned} &\left|2\left(\frac{a_i^i}{2}+\frac{a_i^i}{\tau}\right)\tilde{x}_i^{k+1}+\left(\frac{2a_i^i\gamma}{\tau}+\lambda\right)\mathrm{sgn}(x_i^k)\right|^2\\ &\leq\left(b_i-\langle a^i,y^{k,i-1}\rangle+\left(\frac{2a_i^i\gamma}{\tau}+\lambda\right)\mathrm{sgn}(x_i^k)\right)^2\dots\\ &\qquad+8\lambda\left(\frac{a_i^i}{2}+\frac{a_i^i}{\tau}\right)|x_i^k|, \end{aligned}$$

then set

$$\begin{aligned} x_i^{k+1} &= \tilde{x}_i^{k+1}+\frac{\mathrm{sgn}(x_i^k)}{2\left(\frac{a_i^i}{2}+\frac{a_i^i}{\tau}\right)}\left(\frac{2a_i^i\gamma}{\tau}+\lambda\dots\right.\\ &\left.-\sqrt{\left(b_i-\langle a^i,x^k\rangle+\left(\frac{2a_i^i\gamma}{\tau}+\lambda\right)\mathrm{sgn}(x_i^k)\right)^2+8\lambda\left(\frac{a_i^i}{2}+\frac{a_i^i}{\tau}\right)|x_i^k|}\right),\\ r_i^{k+1} &= -r_i^k. \end{aligned}$$

6.5 Equivalence of iterative methods for linear systems

In what follows, we discuss and demonstrate equivalencies for different iterative methods for solving linear systems. We recall from the previous section that the SOR method (6.11) is equivalent to the Itoh–Abe discrete gradient method [153].

The explicit coordinate descent method [15, 221] is given by

$$\begin{aligned} y^{k,0} &= x^k \\ y^{k,i} &= y^{k,i-1} - \alpha_i [\nabla F(y^{k,i-1})]_i e^i, \\ x^{k+1} &= y^{k,n}, \end{aligned} \tag{6.15}$$

where $\alpha_i > 0$ is the time step. As mentioned in [221], the SOR method is also equivalent to the coordinate descent method with F given by (6.10) and the time step $\alpha_i = \omega / a_i^i$. Hence, in this setting, the Itoh–Abe discrete gradient method is equivalent not only to SOR methods, but to explicit coordinate descent.

It is not surprising that these iterative coordinate methods turn out to be the same, given that the gradient F in (6.10) is linear. Furthermore, these equivalencies extend to discretisations of the ISS flow with J given by (6.12). The resultant Bregman Itoh–Abe scheme for (6.10) is described in Section 6.4.1. We may compare this to a *Bregman linearised coordinate descent* scheme,

$$\begin{aligned} y^{k,0} &= x^k, \quad p^k \in \partial J(x^k) \\ z_i &= \arg\min_y [\nabla F(y^{k,i-1})]_i \cdot y + \frac{a_i^i}{\alpha_i} D_J^{p^k}(y^{k,i-1} + ye^i, y^{k,i-1}), \\ y^{k,i} &= y^{k,i-1} + z_i e^i, \\ x^{k+1} &= y^{k,n}. \end{aligned}$$

One can verify that this scheme is equivalent to (6.15) for the parameters

$$\tau_i = \frac{2\alpha}{(2-\alpha)a_i^i}, \quad \lambda^* = \frac{\lambda}{1 + \frac{\alpha}{2-\alpha}}.$$

6.6 Numerical examples

In this section, we present numerical results for the schemes described in Section 6.4.

6.6.1 Sparse SOR

We construct a matrix $A \in \mathbb{R}^{1024 \times 1024}$ from independent standard (zero mean, unit variance) Gaussian draws, and construct the sparse ground truth x^{true} by choosing 10% of the indices at random determined by uniform draws on the unit interval. We then solve the problem

$$\arg\min_x \frac{1}{2} \|Ax - b\|^2,$$

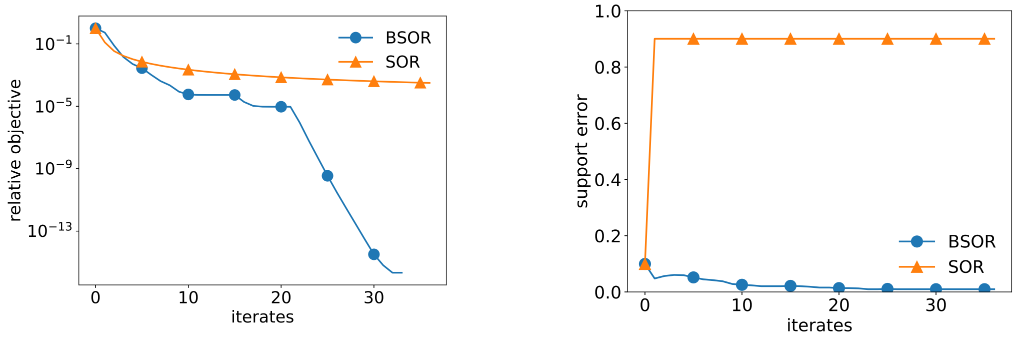


Fig. 6.1 Comparison of SOR and sparse SOR methods, for Gaussian linear system without noise. Left: Convergence rate for relative objective, i.e. $[F(x^k) - F^*]/[F(x^0) - F^*]$. Right: Support error with respect to iterates, i.e. proportion of indices i s.t. $\mathrm{sgn}(x_i^k) = \mathrm{sgn}(x_i^*)$.

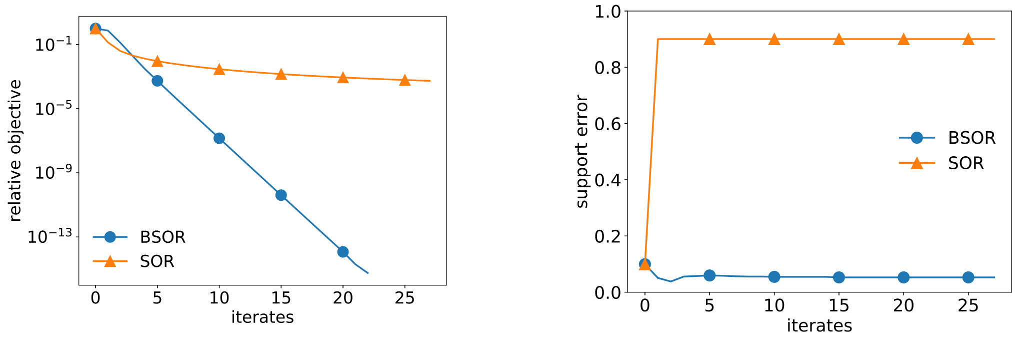


Fig. 6.2 Comparison of SOR and sparse SOR methods, for Gaussian linear system without noise, and binary ground truth. Left: Convergence rate for relative objective. Right: Support error with respect to iterates.

where $b = Ax^{\text{true}}$. We compare the SOR method ($J(x) = \|x\|^2/2$) and the BSOR method ($J(x) = \|x\|^2/2 + \gamma\|x\|_1$), where $\gamma = 1$. We set time steps to $\tau = 2/\mathrm{diag}\,(A)$, corresponding to the Gauss-Seidel method. See Figure 6.1 for the results.

For the same test problem, but where the ground truth is binary, i.e. only takes values 1 or 0, see Figure 6.2.

6.6.2 Sparse, regularised SOR

We construct $A \in \mathbb{R}^{1024\times 1024}$ and x^{true} as in the previous subsection. However, we add noise to the data, i.e. $b = Ax^{\text{true}} + \delta$, where δ is independent Gaussian noise with a standard deviation of $0.1\|Ax^{\text{true}}\|_\infty$. Since the added noise destroys the sparsity structure of $A^{-1}b$, the sparse SOR method fails to improve the convergence rate. The results for $F(x) = \|Ax - b\|^2/2$ are in Figure 6.3.

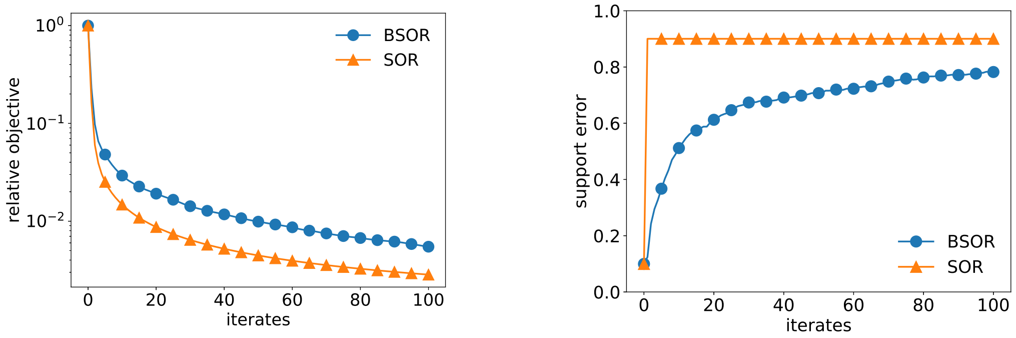


Fig. 6.3 Comparison of SOR and sparse SOR methods, for Gaussian linear system with noise. Left: Convergence rate for relative objective. Right: Support error with respect to iterates.

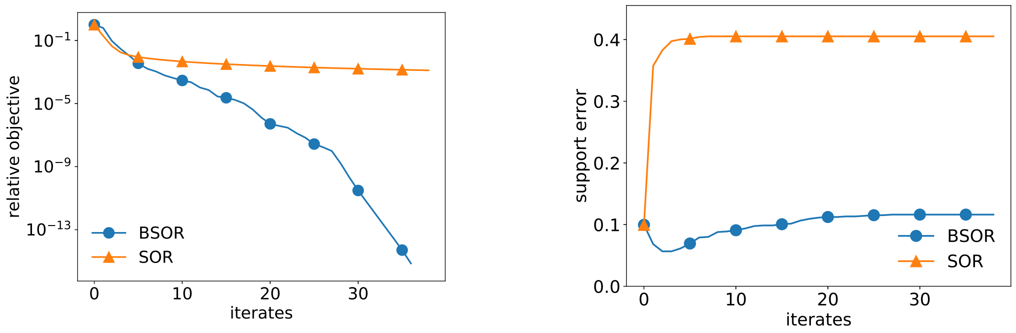


Fig. 6.4 Comparison of SOR and sparse SOR methods, for ℓ^1-regularised linear system with noise. Left: Convergence rate for relative objective. Right: Support error with respect to iterates.

We therefore include regularisation in the objective function of the form

$$F(x) = \frac{1}{2}\|Ax - b\|^2/2 + \lambda\|x\|_1,$$

where $\lambda = 100$, and with initialisation x^0 constructed by random, independent Gaussian draws. The results are visualised in Figure 6.4.

In all of these cases, the sparsity structure when utilised properly leads to significantly faster convergence rates with the BSOR method. We note that while we only consider linear systems, these methods could be implemented for arbitrary problems.

6.7 Conclusion and outlook

In this chapter, we propose to discretise the ISS flow with the Itoh–Abe discrete gradient. The resultant schemes exhibit a dissipative structure (6.5) related to the symmetrised Bregman distance of a function J. This generalises the discrete gradient method for gradient flows, and can be viewed as a discrete gradient version of Bregman iterations. Building on the analysis of Chapter 4, we prove convergence guarantees of the Bregman Itoh–Abe discrete gradient method in the Clarke subdifferential framework.

We consider numerical examples motivated by linear systems and searching for sparse solutions. These results indicate that for sparse reconstructions, popular iterative solvers such as the SOR method can be significantly sped up by incorporating a Bregman step.

Future work is dedicated to proving convergence rates for the Bregman Itoh–Abe methods, and to compare the scheme to related methods such as the sparse Kaczmarz method [141].

Chapter 7

Differentiation for nonsmooth bilevel optimisation

7.1 Introduction

In this chapter, we study bilevel optimisation of nonsmooth variational problems[1]. While in a previous chapter we treated this class of problems as black-box, and employed derivative-free optimisation methods to solve them, we now study methods for differentating the lower-level solution map. Moving from a derivative-free approach to a gradient-based approach can be necessary when the parameter space becomes high-dimensional.

We recall the bilevel problem (1.11) discussed in Chapters 1 and 4. Namely, there is a lower-level variational problem

$$x_\vartheta \in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} F(x,\vartheta), \tag{7.1}$$

and an upper-level problem

$$\min_{\vartheta\in\Omega} E(x_\vartheta,\vartheta), \qquad \text{such that} \qquad x_\vartheta \text{ solves (7.1)}, \tag{7.2}$$

where Ω is an open, connected[2] subset of $\mathbb{R}^m$. For each parameter $\vartheta \in \Omega$, $F(\cdot,\vartheta) : \mathbb{R}^n \to \overline{\mathbb{R}}$ belongs to $\Gamma_0(\mathbb{R}^n)$. We furthermore assume that $E : \mathbb{R}^n \times \Omega \to \mathbb{R}$ is C^1-smooth. We also

[1] I am grateful to Jingwei Liang for many helpful comments and pointers.

[2] The domain Ω can also be more general. However, we assume the parameters ϑ lie in the interior of the domain, so we do not need to treat differentiation along boundaries.

denote by $\overline{E}$ the *bilevel objective function*

$$\overline{E}(\vartheta) := E(x_\vartheta, \vartheta). \tag{7.3}$$

As previously discussed, bilevel optimisation poses several challenges, due to each point evaluation involving the minimisation of a variational function. When the lower-level problem is not strictly convex or nonconvex, the relation $\vartheta \mapsto x_\vartheta$ might not be unique, leading to discontinuities in the upper-level problem, and could require one to treat it as a set-valued optimisation problem. Furthermore, from a practical point of view, bilevel optimisation is computationally intensive due to the typically high cost of solving (7.1) for each parameter choice, and the difficulty in computing gradients of $\vartheta \mapsto x_\vartheta$ (if they exist). Even when x_ϑ is uniquely defined, its dependence on ϑ tends to be highly nonlinear, and, in the case of nonsmooth variational problems, nonsmooth.

7.1.1 Contributions and structure of chapter

In this chapter, we study the first-order behaviour of x_ϑ with respect to ϑ, and the corresponding impact of nonsmoothness of F. Broadly speaking, there are two approaches to computing gradients of $\vartheta \mapsto x(\vartheta)$; implicit differentiation and algorithmic differentiation. In both cases, the nonsmoothness of F needs to be accounted for.

For these purposes, we will use the *partial smoothness* framework, a powerful framework for nonsmooth optimisation analysis. Introduced by Lewis in 2002 [130], it is motivated by the premise that "nonsmoothness pervades optimization, but the way it typically arises is highly structured". We apply this framework to show local piecewise differentiability of the solution map, and based on this, we provide an expression for the Clarke subdifferential of the bilevel objective function, i.e. $\partial^C\overline{E}$. Furthermore, in the setting of algorithmic differentiation, we prove convergence of the algorithmic derivatives to the limiting implicit gradient for various forward-backward type methods under a standard nondegeneracy assumption within the partial smoothness framework.

The rest of the chapter is structured as follows. In Section 7.2, we review the current literature on bilevel optimisation in signal processing. We also review works on partial smoothness. In Section 7.3, we introduce preliminary concepts required for our results, and outline the conditions required to ensure sufficient regularity of F, which in general will be nonsmooth and nonconvex. In Section 7.4, we characterise and prove the piecewise differentiability of $\vartheta \mapsto x_\vartheta$ and the Clarke subdifferential of $\overline{E}$. In Section 7.5, we study algorithmic differentiation of forward-backward type algorithms, including the accelerated version FISTA, and some Bregman proximal gradient variants. Finally, we provide numerical results in Section 7.6.

In order to give full attention to the impact of nonsmoothness of F on the solution mapping, we make some simplifications and do not address other practical and theoretical considerations of bilevel optimisation. The strongest assumption we make is that of strong convexity of $F(\cdot,\vartheta)$, which will ensure that the solution map is well-defined and continuous globally, i.e. for all ϑ. However, all the results could be given in a more general setting, locally for a neighbourhood of parameters. For example, strong convexity could be replaced with *restricted injectivity* [135] or *restricted positive definiteness* [214], which would yield the same results locally, but not necessarily globally. In fact, much of the theory of partial smoothness does not require convexity, e.g. “convexity is *not* the real driving force behind this theory" [130]; “it is worth noting that convexity (and even Clarke regularity) is of no consequence for us" [76]. Furthermore, we do not consider schemes for solving the upper-level problem (7.2) once an acceptable derivative $Dx(\vartheta)$ has been computed, nor do we consider any constraints on the parameter space. We therefore view the computation of $Dx(\vartheta)$ as separate from the actual bilevel optimisation scheme of choice.

7.2 Literature review

In order to contextualise gradient-based approaches to nonsmooth bilevel optimisation, we first need to discuss classical implicit differentiation. Recalling the implicit function theorem Proposition 2.3, we observe that if $F(\cdot,\vartheta)$ were C^2-smooth and strongly convex, then the solution map $x(\vartheta)$ would be the unique solution to the first-order condition $0=\nabla_x F(x,\vartheta)$, where $\nabla_x F$ is C^1-smooth and $\nabla_x^2 F$ is positive-definite by strong convexity. In this case, the implicit function theorem can be applied directly to show that $x(\vartheta)$ is C^1-smooth and

$$Dx(\vartheta) = -(\nabla_x^2 F(x(\vartheta),\vartheta))^{-1} D_\vartheta \nabla_x F(x(\vartheta),\vartheta). \tag{7.4}$$

When F is not C^2-smooth, or even differentiable, there is then the question of what can still be done.

7.2.1 Bilevel problems and smoothed lower-level problems

We review previous works on bilevel problems with nonsmooth lower-level problems, and approaches to gradient-based optimisation.

Kunisch & Pock in [125] consider bilevel optimisation problems of the form

$$\min_{\vartheta \geq 0} \|x(\vartheta) - x^{\dagger}\|^2 \quad \text{s.t.} \quad x(\vartheta) \in \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \frac{1}{p} \sum_{k=1}^{q} \vartheta_k \|A_k x\|_p^p + \frac{1}{2}\|x - f\|^2. \tag{7.5}$$

Here $x^{\dagger}$ is the ground truth, f is the data, and each of the q linear terms A_k are referred to as analysis-based priors, and which we will refer to as filters. In this case, the parameters to be trained can be seen as weights of each filter term. For the theoretical treatment, they consider $p \in \{1,2\}$ which yields nonsmooth and smooth bilevel problems respectively. For their implementations, the nonsmooth terms are smoothed, after which they can apply the classical implicit function theorem. They also trained the parameters for $p = 1/2$, which yields a nonconvex optimisation problem whose gradient blows up at $x = 0$. While $p = 1/2$ leads to theoretical and computational difficulties, they report that it can lead to significant improvements in denoising images, demonstrating the potential of nonconvex regularisation models.

Fehrenbach et al. in [88] propose a bilevel optimisation model for denoising images in cases where the noise is from a known distribution, e.g. Gaussian noise. The lower-level problem is essentially that of (7.5) for $p = 1$. However the upper-level objective function is given as $E(x(\vartheta), \vartheta) = G(f - x(\vartheta))$, where G is a measure of Gaussianity (in the case of Gaussian noise). The idea is to identify parameters such that the residual of the reconstruction, i.e. $f - x$, fits the known noise statistics. This is therefore an unsupervised bilevel problem. To differentiate the solution map, they propose to use smoothened ℓ^1-norms and apply implicit differentiation.

The two first examples presented above deal with learning the optimal coefficients for a collection of regularisation terms. Another important class bilevel problem deals with learning the regularisation terms themselves, namely learning analysis priors. Peyré & Fadili in [177] present the following bilevel problem,

$$\min_{D \in \mathbb{R}^{n,p}} \frac{1}{2} \sum_{k=1}^{q} \|x^{\dagger,k} - x(D, f^k)\|^2, \quad \text{s.t.} \quad x(D, f^k) \in \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \frac{1}{2}\|x - f^k\|^2 + \Gamma(D^* x), \tag{7.6}$$

where $x^{\dagger,k}$, f^k, $k = 1, \ldots, q$ denote a collection of ground truth images and corresponding noisy images, and where Γ is typically the ℓ^1-norm or a smoothened version. *Dictionary learning* problems [113, 145, 169] refer to various approaches of learning a dictionary D such that the lower-level problem in (7.6) yields an optimal reconstruction.

Chen et al. in [52] relate the aforementioned dictionary models above to Markov random field (MRF) models, such as Field of Experts [195], proposed in 2009 by Roth & Black.

MRF models seek to learn image priors that capture the statistics of natural images, and are therefore naturally related to dictionary learning within the realm of bilevel optimisation. Extensions of filter learning applications include recent work by Benning et al. [19] which proposes to learn filters that simultaneously promote desirable signal features and penalise undesirable features. This is modelled as a quotient minimisation problem.

7.2.2 Nonsmooth analysis for bilevel optimisation

In all of the examples of bilevel optimisation discussed above, the lower-level problems are assumed to be smooth, typically by smoothening the ℓ^1-norm, and the bilevel gradients are computed using implicit differentiation. However, there are several works on bilevel optimisation that deal with nonsmoothness in the lower-level problem.

Hintermüller & Wu in [111] propose a bilevel optimisation method to learn the point spread function in blind deconvolution problems. The use of TV regularisation introduces nonsmoothness to the lower-level problem, and an analysis of the regularity properties of the nonsmooth solution map is carried out, using Robinson's framework of strong regularity [186]. They derive Bouligand differentiability (local Lipschitz continuity and directional differentiability [187]) of the solution map, and propose a proximal gradient method whose iterates converge to a Clarke stationary point. Furthermore, under a strict complementarity assumption, local C^1-smoothness of the solution is derived. Although the analysis is not based on partial smoothness, many of the results can be seen as instances of the results we will present in this chapter, including directional differentiability and local C^1-smoothness under the nondegeneracy assumption (ND), which is equivalent to the strict complementarity condition for this problem. However, the results in this chapter covers a great number of other bilevel problems.

7.2.3 Algorithmic differentiation

The aforementioned works mainly consider implicit differentiation. An alternative approach for evaluating derivatives we must consider is algorithmic differentiation [99], also referred to as automatic differentiation. If we view the solution map as defined implicitly via the first-order condition of a variational problem, then implicit differentiation will yield the derivative. However, if we view the solution map as the output of an algorithm that takes the parameters ϑ as an input, then we can compute the derivative by differentiating the algorithm and applying the chain rule. This framework fits naturally for bilevel optimisation in cases where the lower-level problem is solved via proximal splitting algorithms.

There are several recent works that study algorithmic differentiation with connections to bilevel optimisation. Ochs et al. [166, 167] propose a framework for smooth algorithmic differentiation of the solution map, even when the lower-level problem is nonsmooth. This is done by applying iterative methods using Bregman proximal maps, where the Bregman distance generating function is chosen to ensure that each update is differentiable with respect to the parameters, analogous to the use of barrier functions for interior point updates.

Deledalle et al. [68] consider the optimisation of regularisation parameters for inverse problems in the presence of Gaussian noise. Here they extend a framework of unbiased risk estimation to handle nonsmooth variational problems, making use of the fact that their solution map of interest is often Lipschitz continuous, and therefore weakly differentiable. Furthermore, they propose algorithmic differentiation of various iterative proximal splitting algorithms, making use of weak differentiability of the proximal updates and the chain rule for weakly differentiable functions. In their work, as well as in the aforementioned works of Ochs et al., convergence guarantees for the algorithmic derivatives to the implicit differential is left as an open problem.

Deledalle et al. [67] propose a framework for correcting for systematic errors that occur with variational regularisation methods, wherein they employ algorithmic differentiation of proximal methods to compute the Jacobian of the solution mapping. They obtain convergence guarantees for the algorithmic derivatives in [67, Theorem 21] for a class of ℓ^1-regularised problems. In contrast, our convergence results assume a nondegeneracy condition, but cover more general classes of variational problems and algorithmic methods.

Finally we mention the recent work by Bertocchi et al. [21], which considers algorithmic differentiation of iterative proximal methods of variational problems for the optimisation of model parameters as well as algorithmic parameters. This builds on the theory of Combettes & Pesquet [56] which relate deep neural structures to variational problems via proximal mappings.

7.2.4 Bilevel problems in function spaces

This chapter focuses on variational problems in finite-dimensional spaces. This is because, with the exception of a few cases, it is unclear whether and how the partial smoothness framework can be extended to the infinite-dimensional setting. Regardless, there are several important works on the theoretical treatment of bilevel optimisation problems in function spaces.

De los Reyes et al. [65] consider bilevel optimisation problems similar to (7.5) but with the data fidelity term including a linear forward model K, i.e. $||Kx-f||^2/2$ and x defined in a function space. They derive the existence of optimal parameters and outer semicontinuity of the solution mapping, as well as convergence of the smoothed (Huber regularised) bilevel problem to the nonsmooth problem. In [66], De los Reyes et al. propose a semismooth Newton algorithm for solving bilevel problems involving a smoothed TV regulariser.

7.2.5 Partial smoothness

Finally, we discuss some works on partial smoothness of particular relevance to us. Lewis proposed this framework in 2002 [130], seeking to characterise and unify notions of “active" behaviour across various types of nonsmooth optimisation problems. Central to this is the idea of an *active manifold* containing the minimiser, along which the objective function varies smoothly, and such that optimisation algorithms *identify* this manifold after a finite number of iterations. Theoretical results include showing local C^1-smooth behaviour of the solution map under a nondegeneracy assumption, and calculus rules for partly smooth functions, including a chain rule and a sum rule.

Vaiter et al. [214] consider regularised regression problems with nonsmooth regularisers, and show that if the regulariser is partly smooth, one can apply implicit differentiation along the active manifold to compute gradients of their solution map. Furthermore, they consider the case where the nondegeneracy assumption fails, and show that in this case, the set of points where the solution map is nonsmooth has zero Lebesgue measure, assuming that the lower-level objective function is definable in the o-minimality framework. The results in Section 7.4 have connections to the analysis in their paper. In particular, the studies in this chapter is predicated on the implicit differentiability for partly smooth variational problems, and we also consider the differentiability properties of the solution map when the nondegeneracy assumption fails. Our approach differs from the one in [214], i.e. we do not require definability, and we prove piecewise differentiability of $x(\vartheta)$ and explicitly characterise the Clarke subdifferential of $\overline{E}$.

Liang et al. [134, 137, 135, 136] study various iterative proximal splitting algorithms for partly smooth variational problems, proving results such as finite activity identification and consequently local linear convergence. The algorithms include forward–backward type methods, primal dual splitting methods, Douglas–Rachford splitting and ADMM.

We emphasise that there are several additional works of significance on partial smoothness, which this review does not cover. Furthermore, it goes without saying that there is a vast literature on general approaches to sensitivity analysis and parameter-tuning in the setting of nonsmooth, constrained optimisation, which we do not attempt to review for the sake of brevity.

7.3 Preliminary material

In this section, we introduce definitions and concepts that will be used for the study of the solution mapping in the following sections. In particular, we cover concepts of differentiability for piecewise smooth functions, subdifferentially regular functions, and functions defined on Riemannian manifolds respectively.

7.3.1 Piecewise smoothness and semidifferentiability

An important aspect of the solution maps for nonsmooth lower-level problems is that as the solution map transitions from one active manifold to another, it also transitions from one regime of local differentiability to another, thus inducing nonsmoothness in the solution mapping. As we show in Section 7.4.2, the solution map turns out to be piecewise C^1-differentiable in this case.

Definition 7.1 (Piecewise smoothness). *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is* piecewise smooth *on an open set $U \subset \mathbb{R}^n$ if f is continuous on U and for each $x \in U$ there is a finite collection of C^1-smooth functions $f_i, i \in I$ defined on a neighbourhood of x, such that, for some $\varepsilon > 0$ one has $f(y) \in \{f_i(y) \ : \ i \in I\}$ when $|y - x| < \varepsilon$.*

We call the collection of functions $\{f_i \ : \ i \in I(x)\}$ a local representation *of f at x. We call a local representation* minimal *if no proper subset of the collection forms a local representation of f at x.*

It is straightforward to see that piecewise smooth functions are locally Lipschitz continuous.

We present some further results on the regularity of this class of functions. First, we consider a generalisation of differentiability, called *semidifferentiability*. For this, we note that a map is *positively homogeneous* if for all $h \in \mathbb{R}^n$, $t > 0$, one has $\varphi(th) = t\varphi(h)$.

Definition 7.2 (Semidifferentiability). *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is* semidifferentiable *at $x \in \mathbb{R}^n$ if there is a continuous, positively homogeneous mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ such that*

$$f(x+h) = f(x) + \varphi(h) + o(h).$$

The mapping φ is unique, it is called the semiderivative *of f at x, and is denoted by $Df(x;\cdot)$.*

Example 7.3. *The function $f(x) = \|x\|$ is semidifferentiable at 0 with $Df(0;h) = \|h\|$.*

There are many notions of differentiability for nonsmooth functions. However, for locally Lipschitz continuous functions on finite-dimensional spaces, they often coincide. For example, in this case, semidifferentiability coincides with directional differentiability [74, Proposition 2D.1], i.e.

$$Df(x;h) = \lim_{t \downarrow 0} \frac{f(x+th) - f(x)}{t}.$$

See [204], and in particular its Proposition 3.5, for further concepts of differentiability that coincide in this setting.

Proposition 7.4 (Semidifferentiability of piecewise smooth mappings [74, Proposition 2D.8]). *If a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is piecewise smooth on an open set U, then it is semidifferentiable on U, and the semidifferential $Df(x)$ is itself piecewise smooth with a local representation given by $\{\nabla f_i \, : \, i \in I(x)\}$. We refer to these representatives as* local gradient representatives.

Remark 7.5. *Going back to Example 7.3, note that while $f = \|\cdot\|_2$ is semidifferentiable, it is not piecewise smooth, as no finite collection of differentiable functions could form a local representation of f at 0. In contrast, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are both piecewise smooth.*

An important feature of semidifferentiable functions is that they satisfy the chain rule.

Proposition 7.6 (Chain rule). *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^l$ be locally Lipschitz continuous and semidifferentiable at x and $y = f(x)$ respectively. Then their composition $g \circ f : \mathbb{R}^n \to \mathbb{R}^l$ is semidifferentiable at x with semidifferential $D(g \circ f)(x) = Dg(f(x)) \circ Df(x)$.*

Proof. This follows immediately from Proposition 3.5 and Proposition 3.6 in [204]. □

The next result can be verified simply by checking that each condition for piecewise differentiability holds.

Proposition 7.7. *Let $\widetilde{f} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$ be piecewise C^1-differentiable, and denote by $f : \mathbb{R}^n \to \mathbb{R}^l$ its restriction $x \mapsto \widetilde{f}(x, 0)$. Then f is piecewise C^1-differentiable with semiderivative $Df(x;v) = D\widetilde{f}(x, 0; [v, 0]^T)$.*

7.3.2 Generalised differentials and regularity

A requirement for partial smoothness is *subdifferential regularity*. While this always holds for functions in $\Gamma_0(\mathbb{R}^n)$, we consider lower-level objective functions that are nonsmooth and nonconvex. We therefore need to verify that the class of problems we are interested in will be sufficiently regular. In what follows, we first introduce several concept relating to regularity, then present a general form for parametrised variational objective functions, and prove their regularity.

First we define *subgradients* and *regular subgradients* for (nonconvex) functions.

Definition 7.8 (Subdifferentials). *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be an extended function and $x \in \operatorname{dom} f$. The* regular subdifferential $\widehat{\partial} f(x)$ *consists of $v \in \mathbb{R}^n$ for which*

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|).$$

The (general) subdifferential $\partial f(x)$ *consists of $v \in \mathbb{R}^n$ for which*

$$v = \lim_{k\to\infty} v^k, \quad v^k \in \widehat{\partial} f(x^k), \quad x^k \to x, \quad f(x^k) \to f(x).$$

Third, the horizon subdifferential $\partial^\infty f(x)$ *is the set*

$$\{v \,:\, \exists x^k \to x,\, v^k \in \partial f(x^k),\, \lambda_k \downarrow 0 \text{ with } \lambda_k v^k \to v\}.$$

The vectors are called regular subgradients, (general) subgradients, *and* horizon subdifferentials *respectively.*

For convex functions, both the regular and general subdifferentials coincide with the convex subdifferential [192, Proposition 8.12]. These subdifferentials generally do not coincide with the Clarke subdifferential, as the following example shows.

Example 7.9. *Let $f(x) = -|x|$ at $x = 0$. Then the Clarke, regular, and general subdifferentials are given by*

$$\partial^C f(0) = [-1, 1], \quad \widehat{\partial} f(0) = \emptyset, \quad \partial f(0) = \{-1, 1\}.$$

For a further comparison of subdifferentials, see [24], and [192, Theorem 8.49] and its subsequent discussion.

To define regularity of sets and functions, we also need to define *normal spaces* of sets and *epigraphs* of functions.

Definition 7.10 (Normal spaces). *Let $C \subset \mathbb{R}^n$ and $x \in \mathbb{R}^n$. The* regular normal space of C at x, *written $\widehat{N}_C(x)$, is the set of vectors $v \in \mathbb{R}^n$ such that*

$$\langle v, y - x \rangle \leq o(\|y - x\|) \quad \forall y \in C.$$

The (general) normal space of C at x, *written $N_C(x)$, is the set of vectors $v \in \mathbb{R}^n$ such that there are sequences $x^k \to x$ and $v^k \to v$ such that $x^k \in C$ and $v^k \in \widehat{N}_C(x^k)$. These vectors are called* regular normal vectors *and* normal vectors *respectively.*

Definition 7.11 (Horizon cone)**.** *For a set $C \subset \mathbb{R}^n$, the* horizon cone *is the closed cone C^∞ given by*

$$C^\infty = \begin{cases} \{x \,:\, \exists x^k \in C,\ \lambda_k \downarrow 0,\ s.t.\ \lambda_k x^k \to x\}, & \text{if } C \neq \emptyset, \\ \{0\}, & \text{if } C = \emptyset. \end{cases}$$

Proposition 7.12 (Indicator functions)**.** *Let $C \subset \mathbb{R}^n$ be regular at $x \in C$. Then it holds that*

$$\partial \delta_C(x) = N_C(x) = \widehat{N}_C(x) = \widehat{\partial} \delta_C(x). \tag{7.7}$$

Proof. The first and final equalities are straightforward to verify from definition, while the second equality follows from regularity of C at x. □

Definition 7.13 (Epigraph)**.** *The* epigraph *of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the set*

$$\operatorname{epi} f := \{(x, \alpha) \,:\, \alpha \geq f(x)\}.$$

Definition 7.14 (Subdifferential regularity)**.** *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called* (subdifferentially) regular *at $x \in \operatorname{dom} f$ if $\operatorname{epi} f$ is Clarke regular at $(x, f(x))$ as a subset of $\mathbb{R}^n \times \mathbb{R}$.*

We list properties for the lower-level objective function $F(x, \vartheta)$ that are sufficiently general to cover all our problems of interest, and which we will prove are regular.

Assumption 7.15.

(i) The function F can be written as $F(x, \vartheta) = F_0(x, \vartheta) + \delta_C(x)$, where $C \subset \mathbb{R}^n$ is a closed, convex, nonempty set, F_0 is locally Lipschitz continuous for each $(x, \vartheta) \in \operatorname{dom} F$, and $x \mapsto F_0(x, \vartheta)$ is convex for each $\vartheta \in \Omega$.

(ii) The effective domain of F is independent of ϑ, i.e. $\operatorname{dom} F = \operatorname{dom} F(\cdot, \vartheta) \times \Omega$ for any $\vartheta \in \Omega$.

(iii) For each $x \in \operatorname{dom} F(\cdot, \vartheta)$, $\vartheta \mapsto F(x, \vartheta)$ is C^1-smooth.

(iv) The mapping $(x, \vartheta) \mapsto \partial_x F(x, \vartheta)$ is outer semicontinuous on $\operatorname{dom} F$.

(v) The mapping $(x, \vartheta) \mapsto D_\vartheta F(x, \vartheta)$ is continuous on $\operatorname{dom} F$.

Theorem 7.16. *Let $F : \mathbb{R}^n \times \Omega \to \overline{\mathbb{R}}$ be a function that satisfies Assumption 7.15. Then F is regular at all $(x^*, \vartheta^*) \in \operatorname{dom} F$ and*

$$\partial F(x^*, \vartheta^*) = \partial_x F(x^*, \vartheta^*) \times \{D_\vartheta F(x^*, \vartheta^*)\} \neq \emptyset. \tag{7.8}$$

Proof. We first show regularity of the second term of F. Since $(x, \vartheta) \mapsto \delta_C(x)$ is a proper, lower semicontinuous, convex function, regularity of δ_C on C follows from [192, Proposition 8.12] and [192, Corollary 8.11].

Next we show regularity of F_0. By [192, Corollary 8.11], this holds if and only if

$$\partial^\infty F_0(x^*, \vartheta^*) = \widehat{\partial} F_0(x^*, \vartheta^*)^\infty, \qquad \partial F_0(x^*, \vartheta^*) = \widehat{\partial} F_0(x^*, \vartheta^*).$$

Since F_0 is locally Lipschitz continuous at (x^*, ϑ^*), it follows from [192, Theorem 9.13] that $\partial^\infty F_0(x^*, \vartheta^*) = \widehat{\partial} F_0(x^*, \vartheta^*)^\infty = \{0\}$. Since $\partial F_0(x^*, \vartheta^*) \supset \widehat{\partial} F_0(x^*, \vartheta^*)$, it remains to show that $\partial F_0(x^*, \vartheta^*) \subset \widehat{\partial} F_0(x^*, \vartheta^*)$. Suppose $v = [v_x, v_\vartheta]^T \in \partial F_0(x^*, \vartheta^*)$, so there are sequences $(x^k, \vartheta^k) \to (x^*, \vartheta^*)$ and $v^k \in \widehat{\partial} F_0(x^k, \vartheta^k)$ such that $v^k = [v_x^k, v_\vartheta^k]^T \to v$.

As F_0 is convex in the first argument and continuously differentiable in the second, the equalities

$$\partial_x F_0(x, \vartheta) = \widehat{\partial}_x F_0(x, \vartheta), \quad \partial_\vartheta F_0(x, \vartheta) = \widehat{\partial}_\vartheta F_0(x, \vartheta) = \{D_\vartheta F_0(x, \vartheta)\}.$$

follow from [192, Theorem 9.18] and [192, Proposition 8.12]. Furthermore, by the previous equalities and [192, Corollary 10.11], we have

$$\widehat{\partial} F_0(x, \vartheta) \subset \partial F_0(x, \vartheta) \times \partial_\vartheta F_0(x, \vartheta) = \partial_x F_0(x, \vartheta) \times \{D_\vartheta F_0(x, \vartheta)\}.$$

Therefore, $v^k = [v_x^k, D_\vartheta F_0(x^k, \vartheta^k)]^T$, where $v_x^k \in \partial_x F_0(x^k, \vartheta^k)$. By Assumption 7.15, $\partial_x F_0$ is outer semicontinuous and $D_\vartheta F_0$ is continuous, so $v \in \partial_x F_0(x^*, \vartheta^*) \times \{D_\vartheta F_0(x^*, \vartheta^*)\}$.

Finally, we show that $v \in \widehat{\partial} F_0(x^*, \vartheta^*)$. Suppose $(h^k, r^k) \to (0, 0)$. Then

$$\begin{aligned} &F_0(x^* + h^k, \vartheta^* + r^k) - F_0(x^*, \vartheta^*) \\ &\qquad = F(x^* + h^k, \vartheta^* + r^k) - F_0(x^* + h^k, \vartheta^*) + F_0(x^* + h^k, \vartheta^*) - F_0(x^*, \vartheta^*) \\ &\qquad \geq \langle v_x, h^k \rangle + \langle D_\vartheta F_0(x^* + h^k, \vartheta^*), r^k \rangle + o(\|r^k\|) \\ &\qquad = \langle v_x, h^k \rangle + \langle D_\vartheta F_0(x^*, \vartheta^*), r^k \rangle + o(\|r^k\|), \end{aligned}$$

where the inequality follows from convexity of F in the first argument and the final equality is due to continuity of $D_\vartheta F_0$. Thus F_0 is regular at (x^*, ϑ^*).

Since δ_C and F_0 are regular, and $\partial^\infty F_0(x^*, \vartheta^*) = \{0\}$, [192, Corollary 10.9] implies that F is regular at (x^*, ϑ^*) and that the subdifferential is given by (7.8). This concludes the proof. □

7.3.3 Riemannian geometry

We now introduce concepts relating to functions defined on Riemannian manifolds. We say that $\mathcal{M} \subset \mathbb{R}^n$ is a *l-dimensional C^2-smooth submanifold around $x \in \mathcal{M}$* if there is a C^2-smooth function $G : N \to \mathbb{R}^l$ such that

$$\mathcal{M} \cap N = \{y \in N \;:\; G(y) = 0\} \cap N,$$

and where the matrix $\nabla G(y)$ is surjective for all $y \in N$ [63]. For brevity, we will from hereon refer to these as smooth manifolds. The *tangent space $T_x\mathcal{M}$* and *normal space $N_x\mathcal{M}$* are given by

$$T_x\mathcal{M} = \ker \nabla G(x), \qquad N_x\mathcal{M} = (\ker \nabla G(x))^{\perp}.$$

Definition 7.17 (Smooth representative). *Given a set $\mathcal{M} \subset \mathbb{R}^n$ and a function $f : \mathcal{M} \mapsto \overline{\mathbb{R}}$, we call f* smooth along $\mathcal{M}$ at $x \in \mathcal{M}$ *if there is a neighbourhood N_x of x in $\mathbb{R}^n$, and a smooth function $g : N_x \to \overline{\mathbb{R}}$ such that $g(y) = f(y)$ for all $y \in \mathcal{M} \cap N_x$. We call g a* smooth representative of f around x.

We can now define the *Riemannian gradient* and *Riemannian Hessian*, which will be central to implicit differentiation of nonsmooth functions. Throughout, we denote by $\mathcal{M} \subset \mathbb{R}^n$ a smooth manifold.

Definition 7.18 (Riemannian gradient). *Let $f : \mathcal{M} \to \mathbb{R}$ be smooth at $x \in \mathcal{M}$ and denote by g any smooth representative of f at x. Then the* Riemannian gradient *of f at x is given by*

$$\nabla_{\mathcal{M}} f(x) := P_{T_x\mathcal{M}} \nabla g(x),$$

where $P_{T_x\mathcal{M}}$ is the projection operator onto the tangent space.

The following result can be found in [63, Proposition 9 & Proposition 12].

Proposition 7.19. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a regular function at $x \in \mathcal{M}$ such that $\partial f(x) \neq \emptyset$, and suppose f is smooth along $\mathcal{M}$ at x. Then the Riemannian gradient of f is independent of the choice of smooth representative of f, and furthermore,*

$$\nabla_{\mathcal{M}} f(x) = P_{T_x\mathcal{M}} \partial f(x).$$

Definition 7.20 (Riemannian Hessian). *Let $f : \mathcal{M} \to \mathbb{R}$ be C^2-smooth at $x \in \mathcal{M}$. The* Riemannian Hessian *of f at x is the symmetric, linear mapping from $T_x\mathcal{M}$ to itself defined as*

$$\langle u, \nabla^2_{\mathcal{M}} f(x) u \rangle := \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(P_{T_x\mathcal{M}}(x + tu))|_{t=0}, \quad \forall u \in T_x\mathcal{M},$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product.

The Riemannian Hessian has an alternative expression, for which we introduce the *Weingarten map* of $\mathcal{M}$ at x. Given a normal vector $v \in N_x\mathcal{M}$, the Weingarten map is the symmetric, linear operator $\mathfrak{W}_x(\cdot, v) : T_x\mathcal{M} \to T_x\mathcal{M}$, given by [51, Proposition II.2.1]

$$\mathfrak{W}_x(u,v) = -P_{T_x\mathcal{M}} dV[u], \quad u \in T_x\mathcal{M}.$$

Here V is any extension of v to a vector field on the normal bundle of $\mathcal{M}$ embedded in $\mathbb{R}^n$, i.e. $\{(x,w) \,:\, x \in \mathbb{R}^n,\ w \in N_x\mathcal{M}\} \subset \mathbb{R}^n \times \mathbb{R}^n$, and dV is the derivative of V under the standard (i.e. Euclidean) connection d on $\mathbb{R}^n$ [51]. For further details on the Weingarten map, also known as the second fundamental form, we refer the reader to [3, 51].

Denote by g any smooth representative of f at x. Then the Riemannian Hessian can alternatively be written as

$$\nabla^2_{\mathcal{M}} f(x)v = P_{T_x\mathcal{M}} \nabla^2 g(x) P_{T_x\mathcal{M}} v + \mathfrak{W}_x(v, -P_{N_x\mathcal{M}} \nabla g(x)), \quad \forall v \in T_x\mathcal{M}.$$

Furthermore, if $\mathcal{M}$ is an affine or linear manifold near x, then $\mathfrak{W}_x(v, -P_{N_x\mathcal{M}} \nabla g(x))$ vanishes, so the Riemannian Hessian simplifies to $\nabla^2_{\mathcal{M}} f(x) = P_{T_x\mathcal{M}} \nabla^2 g(x) P_{T_x\mathcal{M}}$ [214].

7.4 Partial smoothness and implicit differentiation

In what follows, we will define partial smoothness, present our class of lower-level objective functions, and derive properties of the corresponding solution map $\vartheta \mapsto x(\vartheta)$.

For a convex set $C \subset \mathbb{R}^n$, its *affine hull*, denoted by $\operatorname{aff} C$, is the smallest affine set that contains C. We denote by $\operatorname{par} C$ the subspace parallel to $\operatorname{aff} C$.

Definition 7.21 (Relative interior and boundary). *For a convex set $C \subset \mathbb{R}^n$, the* relative interior $\operatorname{ri} C$ *is the interior of C relative to its affine hull. The* relative boundary $\operatorname{rbd} C$ *is given by* $\operatorname{rbd} C := \operatorname{cl} C \setminus \operatorname{ri} C$.

We are now ready to define partial smoothnes.

Definition 7.22 (Partial smoothness). *Let $\mathcal{M}$ a smooth manifold in $\mathbb{R}^n$. The function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* partly smooth at $x \in \mathcal{M}$ relative to $\mathcal{M}$ *if the following properties hold.*

(i) (restricted smoothness) *f restricted to $\mathcal{M}$ is C^2-smooth around x.*

(ii) (subgradient continuity) *The subdifferential ∂f is continuous at x relative to $\mathcal{M}$.*

(iii) (normal sharpness) $\operatorname{par}\big(\partial f(x)\big) = N_{\mathcal{M}}(x)$.

(iv) (regularity) *f is regular at x and $\partial f(x) \neq \emptyset$.*

We summarise the assumptions we make on $F : \mathbb{R}^n \times \Omega \to \overline{\mathbb{R}}$ as follows.

Assumption 7.23.

(A1) For each ϑ, $F(\cdot, \vartheta)$ is in $\Gamma_0(\mathbb{R}^n)$ and is μ-convex for some $\mu > 0$ independent of ϑ.

(A2) For each $x \in \mathbb{R}^n$, there is a neighbourhood $N_x \ni x$ that can be partitioned into a finite number of smooth manifolds $\mathfrak{M}_x = \{\mathcal{M}_i : i = 1, \dots, N\}$ such that for each $y \in \mathcal{M}_i \cap N_x$, and $\vartheta \in \Omega$, F is partly smooth at (y, ϑ) relative to $\mathcal{M}_i \times \mathbb{R}^m$.

The two next assumptions ensure sufficient regularity at the boundary of the manifolds $\mathcal{M} \in \mathfrak{M}_x$.

(A3) For each $x \in \mathbb{R}^n$ and $\mathcal{M} \in \mathfrak{M}_x$, if $\operatorname{cl}\mathcal{M} \setminus \mathcal{M} \neq \emptyset$, then there is a smooth manifold $\widetilde{\mathcal{M}} \subset \mathbb{R}^n$ such that

$$\operatorname{cl}\mathcal{M} \subset \widetilde{\mathcal{M}} \quad \textit{and} \quad T_y\mathcal{M} = T_y\widetilde{\mathcal{M}} \quad \forall y \in \mathcal{M}.$$

(A4) For each $x \in \mathbb{R}^n$ and $\mathcal{M} \in \mathfrak{M}_x$, the limit

$$\lim_{\mathcal{M} \ni y^k \to y \in \operatorname{cl}\mathcal{M}, \vartheta^k \to \vartheta \in \Omega} \partial_x F(y^k, \vartheta^k)$$

is well-defined.

Remark 7.24. *By the transversality embedding assumption and chain rule for partly smooth functions [130, Assumption 5.1 & Theorem 4.2], Assumption 7.23 (A2) implies that F_ϑ is partly smooth at x relative to $\mathcal{M}$ for all ϑ.*

The concept of partitioning a neighbourhood of x into manifolds as in (A3) *is not new in the context of partial smoothness, consider e.g. the framework of mirror-stratifiability in [86]. However, mirror-stratifiable functions are more restrictive than our assumptions, as they also require a duality pairing with the convex conjugate F^*.*

Since F is strongly convex in the first argument, there is a well-defined solution mapping,

$$x(\vartheta) := \operatorname*{arg\,min}_{x \in \mathbb{R}^n} F(x, \vartheta). \tag{7.9}$$

It is straightforward to show that the mapping is continuous.

Proposition 7.25. *Let the function F satisfy Assumption 7.23. Then the functions $\vartheta \mapsto x(\vartheta)$ and $\vartheta \mapsto F(x(\vartheta), \vartheta)$ are continuous.*

Proof. Let $\varepsilon > 0$, $\vartheta^* \in \Omega$ and write $x^* = x(\vartheta^*)$. By the subgradient continuity of F with respect to ϑ, there is $\delta > 0$ such that for all $\vartheta \in \Omega$ with $\|\vartheta - \vartheta^*\| < \delta$, we have $\operatorname{dist}\left(\partial_x F(x^*, \vartheta^*), \partial_x F(x^*, \vartheta)\right) < \mu\varepsilon$. Since $0 \in \partial_x F(x^*, \vartheta^*)$, this means that there is $p \in \partial_x F(x^*, \vartheta)$ such that $\|p\| < \mu\varepsilon$. Finally, by strong convexity of $F(\cdot, \vartheta^*)$, we have

$$\|x^* - x(\vartheta)\| \leq \frac{1}{\mu}\|p\| < \varepsilon.$$

This concludes the proof for $x(\cdot)$.

We prove continuity of $\vartheta \mapsto F(x(\vartheta), \vartheta)$ by contradiction. Suppose there is $\varepsilon > 0$ and $\vartheta_k \to \vartheta^*$ such that $|F(x(\vartheta_k), \vartheta_k) - F(x^*, \vartheta^*)| \geq \varepsilon$ for all k. By continuity of $x(\cdot)$ and lower semicontinuity of F,

$$F(x^*, \vartheta^*) \leq \liminf_{k \to \infty} F(x(\vartheta_k), \vartheta_k) \quad \Longrightarrow \quad F(x(\vartheta_k), \vartheta_k) \geq F(x^*, \vartheta^*) + \varepsilon \;\text{ for } k \geq K.$$

However, Assumption 7.23 implies that $\vartheta \mapsto F(x, \vartheta)$ is continuous. Hence there is k such that $F(x^*, \vartheta_k) < F(x(\vartheta_k), \vartheta_k)$, which contradicts $x(\vartheta_k)$ being a minimiser. Thus $\vartheta \mapsto F(x(\vartheta), \vartheta)$ is continuous. □

7.4.1 Implicit differentiation on the manifold

One of the primary motivations for the framework of partial smoothness by Lewis in [130] was to give conditions under which the minimiser behaves stably with respect to perturbations. For this to be the case, a *nondegeneracy condition* needs to hold. This condition holds for a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at $x \in \mathbb{R}^n$ if

$$0 \in \operatorname{ri}\left(\partial f(x)\right). \tag{ND}$$

A point $x \in \mathbb{R}^n$ is said to be a *strong critical point* of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ relative to a set $\mathcal{M} \subset \mathbb{R}^n$ if (ND) holds for f at x, and f restricted to $\mathcal{M}$ grows quadratically near x. Since our function F is strongly convex with respect to x, we take quadratic growth for granted while discussing minimisation conditions.

Lemma 7.26. *Let $F : \mathbb{R}^n \times \Omega \to \overline{\mathbb{R}}$ be a function that satisfies Assumption 7.23, and which is partly smooth at (x^*, ϑ^*) relative to $\mathcal{M} \times \mathbb{R}^m$ for some smooth manifold $\mathcal{M}$. If* (ND) *holds for F at (x^*, ϑ^*), then there is a neighbourhood N_{ϑ^*} of ϑ^* such that $x(N_{\vartheta^*}) \subset \mathcal{M}$ and $x(\cdot)$ is C^1-smooth on N_{ϑ^*}, with differential*

$$Dx(\vartheta) = -(\nabla^2_{\mathcal{M}} F(x(\vartheta), \vartheta))^{\dagger} D_{\vartheta} \nabla_{\mathcal{M}} F(x(\vartheta), \vartheta). \tag{7.10}$$

Remark 7.27. *Note that the expression for the solution map's differential* (7.10) *derives from implicit differentiation on the active manifold $\mathcal{M}$, and reduces to the classical implicit derivative* (7.4) *when $\mathcal{M} = \mathbb{R}^n$.*

Proof. We first show that $x(\vartheta) \in \mathcal{M}$ for all ϑ sufficiently close to ϑ^*. Write $f(x) := F(x, \vartheta^*)$ for shorthand. By [76, Theorem 4.7], there is $\delta > 0$ such that

$$(\operatorname{gph}\partial f) \cap N = (\operatorname{gph}\partial(f + \delta_{\mathcal{M}})) \cap N, \tag{7.11}$$

for $N = \{x \in B_\delta(x(\vartheta^*)) \ : \ |f(x) - f(x(\vartheta))| < \delta\} \times B_\delta(0)$.

By Proposition 7.25 and partial smoothness, the four mappings $x(\cdot)$, $F(x(\cdot), \cdot)$, $F(x, \cdot)$, and $\partial F(x, \cdot)$ are continuous with respect to ϑ for all x. Therefore, we can choose $\varepsilon > 0$ such that for all $\vartheta \in B_\varepsilon(\vartheta^*)$, one has $x(\vartheta) \in B_\delta(x(\vartheta^*))$, $F(x(\vartheta), \vartheta^*) \in B_\delta(f(x(\vartheta^*)))$, and there is $p \in \partial f(x(\vartheta)) \cap B_\delta(0)$. Therefore by (7.11), $(x(\vartheta), p) \in \operatorname{gph}\partial(f + \delta_{\mathcal{M}})$. As $\|p\| < \infty$, this can only be the case if $x(\vartheta) \in \mathcal{M}$. This concludes the first part.

Local C^1-differentiability of $x(\vartheta)$ is proven in [130, Theorem 5.7]. The expression for Dx is provided in [214, Theorem 1] in the case where $F(x, \vartheta) = g(x, \vartheta) + J(x)$, where g is smooth and J is partly smooth. However, their theorem and corresponding proof are directly applicable to our setting, given that F is C^2-smooth when restricted to $\mathcal{M}$. □

7.4.2 Piecewise smoothness of the solution map

Lemma 7.26 shows that in the nondegenerate case, the solution map is locally continuously differentiable. This begs the question of how likely (ND) is to hold for any given parameter choice.

The good news is that the nondegeneracy condition is a "generic" property. To be more specific, Drusvyatskiy et al. [76] proved that for a lower semicontinuous, semialgebraic function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and the "perturbed" functions $f_v(x) := f(x) - \langle x, v\rangle$, there is a set of full measure in $\mathbb{R}^n$, S, such that for all $v \in S$, all the minimisers of f_v are strong critical points.

However, this is not to say that (ND) holds at $x(\vartheta)$ for almost all $\vartheta \in \Omega$. One can easily devise a lower-level objective function F such that (ND) fails for all parameter choices. Furthermore, even when the condition holds locally, while searching for optimal parameters for the bilevel problem, one can expect the updated solution to move across different manifolds. For these reasons, it is important to characterise the first-order behaviour of $x(\cdot)$ at points of nonsmoothness.

As mentioned in the literature review, in [214] Vaiter et al. addressed this issue, proving that for a class of definable functions in an o-minimal structure, the set of parameters for which $x(\vartheta)$ is nonsmooth has Lebesgue measure zero. We approach the issue from a different perspective, seeking to explicitly characterise the subdifferential of the bilevel objective function.

For these purposes, we define the functions

$$\begin{aligned} \widetilde{F}(x,\vartheta,p) &:= F(x,\vartheta) - \langle p,x\rangle \\ \widetilde{x}(\vartheta,p) &:= \underset{x\in\mathbb{R}^n}{\arg\min}\,\widetilde{F}(x,\vartheta,p). \end{aligned} \tag{7.12}$$

Since $\widetilde{F}$ is strongly convex in the first argument, the solution mapping $\widetilde{x}$ is also well-defined. Furthermore, it is uniquely defined via the optimality condiftion

$$p \in \partial_x F(\widetilde{x}(\vartheta),\vartheta).$$

In fact, one can verify that if Assumption 7.23 holds for F, then it also holds for $\widetilde{F}$ treating (ϑ,p) as the parameters.

For $\vartheta^* \in \Omega$ and $\varepsilon > 0$, consider the set

$$S_\varepsilon(\vartheta^*) := \widetilde{x}(\vartheta^*, B_\varepsilon(0)).$$

By Proposition 7.25 and Assumption 7.23 *(iii)* and by making ε sufficiently small, $S_\varepsilon(\vartheta^*)$ is contained in the union of a finite number of manifolds, indexed by $I_\varepsilon(\vartheta^*)$,

$$\mathcal{M}_i, \quad i \in I_\varepsilon(\vartheta^*), \tag{7.13}$$

such that F is partly smooth relative to each $\mathcal{M}_i$. Since $\varepsilon \mapsto S_\varepsilon(\vartheta^*)$ is decreasing, i.e. $S_\varepsilon(\vartheta^*) \subset S_{\varepsilon'}(\vartheta^*)$ if $\varepsilon < \varepsilon'$, the index set $\varepsilon \mapsto I_\varepsilon(\vartheta^*)$ is also decreasing, and we can define

$$I(\vartheta^*) := \liminf_{\varepsilon\downarrow 0} I_\varepsilon(\vartheta^*).$$

As the following proposition shows, the set $I(\vartheta^*)$ indexes all the manifolds that the solution map $x(\cdot)$ can move to near ϑ^*, and thereby gauges the degeneracy of a minimiser.

Lemma 7.28. *For any $\vartheta^* \in \Omega$, $I(\vartheta^*)$ is finite and nonempty. Furthermore, it is single-valued if and only if* (ND) *holds for F at (x^*,ϑ^*), where $x^* := x(\vartheta^*)$.*

Proof. We showed above that $I_\varepsilon(\vartheta^*)$ is finite for some $\varepsilon > 0$, which implies that $I(\vartheta^*)$ is finite as well. Suppose $\widetilde{F}$ is partly smooth at (x^*,ϑ^*) along $\mathcal{M} \times \mathbb{R}^m$. As $x(\vartheta^*) \in S_\varepsilon(\vartheta^*)$ for all $\varepsilon > 0$, $I(\vartheta^*)$ is nonempty.

If (ND) holds for F at ϑ^*, then by Lemma 7.26, $I(\vartheta^*)$ is single-valued. It therefore remains to show that if (ND) does not hold, then there is $x^k \to x^*$ and $p^k \in \partial F(x^k, \vartheta^*)$ such that $p^k \to 0$ and $x^k \notin \mathcal{M}$ for sufficiently large k.

Choose $q \in \operatorname{ri} \partial F(x^*, \vartheta^*)$. Since $0 \in \partial F(x^*, \vartheta^*)$, by normal sharpness $-q \in N_{x^*}\mathcal{M}$. Define the sequences

$$p^k := -\frac{q}{k}, \quad x^k := \widetilde{x}(\vartheta^*, p^k), \quad d^k := \frac{x^k - x^*}{\|x^k - x^*\|}.$$

The last sequence is well-defined since $p^k \notin \partial F(x^*, \vartheta^*)$ which implies that $x^k \neq x^*$ for all k.

We will prove by contradiction that eventually $x^k \notin \mathcal{M}$. Suppose there is a subsequence (without relabelling) such that $x^k \in \mathcal{M}$ for all k. We denote by $P_{\|}, P_{\perp} : \mathbb{R}^n \to \mathbb{R}^n$ the orthogonal projections onto $T_{x^*}\mathcal{M}$ and $N_{x^*}\mathcal{M}$ respectively. Write $d^k_{\|} := P_{\|} d^k$ and $d^k_{\perp} := P_{\perp} d^k$. Since $p^k \in N_{x^*}\mathcal{M}$, one has

$$0 = \langle p^k, d^k \rangle + \langle p^k, d^k_{\|} - d^k \rangle \geq \mu \|x^k - x^*\| - \|p^k\| \|d^k_{\perp}\|$$

where the inequality follows from strong convexity. Since $\|p^k\| \to 0$, showing that $\|d^k_{\perp}\| = O(\|x^k - x\|)$ will give us a contradiction.

Let $G : \mathbb{R}^n \to \mathbb{R}^l$ be a C^2-smooth function such that $\mathcal{M}$ is locally represented by $\{x : G(x) = 0\}$ around x^*. Then $T_{x^*}\mathcal{M} = \ker \nabla G(x^*)$. We consider the second-order expansion of G around x^* [22, Proposition A.23],

$$G(y) = G(x^*) + \nabla G(x^*)(y - x^*) + \frac{1}{2}(y - x^*)^T \nabla^2 G(x^*)(y - x^*) + o(\|y - x^*\|^2)$$

Plugging in x^k for y and dividing through by $\|x^k - x^*\|$, we get

$$\nabla G(x^*) d^k = -\frac{1}{2}(x^k - x^*)^T \nabla^2 G(x^*) d^k + o(\|x^k - x^*\|) = O(\|x^k - x^*\|).$$

It remains to show that there exists $c > 0$ such that $\|d^k_{\perp}\| \leq c \|\nabla G(x^*) d^k\|$. Since $d^k_{\perp} \in (\ker \nabla G(x^*))^{\perp}$, we have

$$d^k_{\perp} = (\nabla G(x^*))^{\dagger} \nabla G(x^*) d^k, \quad \text{implying that} \quad \|d^k_{\perp}\| \leq \|(\nabla G(x^*))^{\dagger}\| \|\nabla G(x^*) d^k\|.$$

This concludes the proof. □

We now proceed to state and prove the main result of this section, namely the piecewise smoothness of the solution map $x(\cdot)$.

Theorem 7.29. *Let $F : \mathbb{R}^n \times \Omega \to \overline{\mathbb{R}}$ satisfy Assumption 7.23. Then for all $\vartheta^* \in \mathbb{R}^n$, the solution mapping $x(\cdot)$ is piecewise C^1-differentiable at ϑ^*, and its semidifferential $Dx(\vartheta^*)$ is locally represented by*

$$\{-(\nabla^2_{\mathcal{M}_i} F(x(\vartheta^*),\vartheta^*))^\dagger D_\vartheta \nabla_{\mathcal{M}_i} F(x(\vartheta^*),\vartheta^*) \,:\, i \in I(\vartheta^*)\}.$$

Proof. We will show that $\widetilde{x}$ is piecewise C^1-smooth, and invoke Proposition 7.7 to conclude that $x(\cdot) \equiv \widetilde{x}(\cdot,0)$ is as well.

Denote as usual $x^* = x(\vartheta^*)$. We first show that for a neighbourhood $N_{(\vartheta^*,0)}$ of $(\vartheta^*,0)$, $\widetilde{x}(N_{(\vartheta^*,0)}) \subset \cup_{i\in I(\vartheta^*)}\mathcal{M}_i$. Suppose for contradiction that this is not the case, i.e. there is $\mathcal{M} \in \mathfrak{M}_{x^*} \setminus \{\mathcal{M}_i \,:\, i \in I(\vartheta^*)\}$ and $(\vartheta^k,p^k) \to (\vartheta^*,0)$ such that $x^k := \widetilde{x}(\vartheta^k,p^k) \in \mathcal{M}$ for all k. By Assumption 7.23, $0 \in \lim_{k\to\infty} \partial_x F(x^k,\vartheta^k) = \lim_{k\to\infty} \partial_x F(x^k,\vartheta^*)$. Therefore, for any $\varepsilon > 0$, there is $p \in \mathbb{R}^n$ with $\|p\| < \varepsilon$ and $k \in \mathbb{N}$ such that $x^k = \widetilde{x}(\vartheta^*,p) \in S_\varepsilon(x^*)$. But then $\mathcal{M} \in \{\mathcal{M}_i \,:\, i \in I_\varepsilon(\vartheta^*)\}$ for all $\varepsilon > 0$ which is a contradiction.

We fix $i \in I(\vartheta^*)$ and consider $\mathcal{M}_i$. Since $x^* \in \operatorname{cl}\mathcal{M}_i$, by Assumption 7.23 there is an extension of $\mathcal{M}_i$, $\widetilde{\mathcal{M}_i}$ such that $x^* \in \widetilde{\mathcal{M}_i}$. Let G be a smooth representative of $\widetilde{F}$ on $\mathcal{M}_i \times \mathbb{R}^n \times \mathbb{R}^n$, and consider the function

$$g(x,\vartheta,p) := G(x,\vartheta,p) + \delta_{\widetilde{\mathcal{M}_i}}(x).$$

We will show that for a neighbourhood of $(\vartheta^*,0)$, the solution map

$$y^i(\vartheta,p) := \operatorname*{arg\,min}_{y\in N} g(y,\vartheta,p)$$

is well-defined and C^1-smooth, and furthermore that $y^i(\vartheta,p) = \widetilde{x}(\vartheta,p)$ whenever $\widetilde{x}(\vartheta,p) \in \mathcal{M}_i$.

To derive well-definedness and local differentiability of $y^i(\vartheta,p)$, it is sufficient to show that the conditions of [130, Theorem 5.7] hold, namely that g is partly smooth at $(x^*,\vartheta^*,0)$ relative to $\widetilde{\mathcal{M}_i} \times \mathbb{R}^n \times \mathbb{R}^n$ and that x^* is a strong critical point of $g(\cdot,\vartheta^*,0)$ relative to $\widetilde{\mathcal{M}_i}$. Partial smoothness is immediate, since G is smooth and indicator functions of smooth manifolds are partly smooth relative to the manifold.

By the definition of $I(\vartheta^*)$ there is $p^k \to 0$ such that $x^k := \widetilde{x}(\vartheta^*,p^k)$ is a strong local minimiser of F relative to $\mathcal{M}_i$ for each k. Since x^k is also a strong local minimiser of $G(\cdot,\vartheta^*,p^k)$ relative to $\widetilde{\mathcal{M}_i}$, by C^2-smoothness of G, x^* is a strong local minimiser of G relative to $\widetilde{\mathcal{M}_i}$, going via [214, Lemma 4]. Furthermore,

$$\partial_x g(x^*,\vartheta^*,0) = \nabla_x G(x^*,\vartheta^*,0) + N_{x^*}\widetilde{\mathcal{M}_i} = \operatorname{ri}\partial_x g(x^*,\vartheta^*,0),$$

so x^* is a strong critical point of $g(\cdot,\vartheta^*,0)$ relative to $\widetilde{\mathcal{M}_i}$. Hence we can apply [130, Theorem 5.7] to ensure that $y^i(\vartheta,p)$ is well-defined for a neighbourhood $N_{(\vartheta^*,0)}$ around $(\vartheta^*,0)$.

Finally, suppose for $(\vartheta,p)\in N_{(\vartheta^*,0)}$ that $\widetilde{x}(\vartheta,p)\in\mathcal{M}_i$. Then $\widetilde{x}(\vartheta,p)$ is a strong local minimiser of $\widetilde{F}(\cdot,\vartheta,p)$ relative to $\mathcal{M}_i$, and therefore also a strong local minimiser for g. Hence by uniqueness of $y^i(\vartheta,p)$, $y^i(\vartheta,p)=\widetilde{x}(\vartheta,p)$. This concludes the proof. □

As the proof to Theorem 7.29 shows, $\widetilde{x}(\vartheta,p)$ is piecewise smooth with *minimal* local representation $\{y^i(\vartheta,p)\ :\ i\in I(\vartheta^*)\}$. By contrast, $\{y^i(\vartheta,0)\ :\ i\in I(\vartheta^*)\}$ form a local representation of $x(\vartheta)$ but not necessarily a minimal one. However, we will see in the following corollary that whether it is a minimal representation for $x(\cdot)$ or not does not matter for the Clarke subdifferential of the bilevel objective function.

We denote the bilevel functions accordingly.

$$\widetilde{E}(\vartheta,p):=E(\widetilde{x}(\vartheta,p),\vartheta),\quad \widetilde{E}^i(\vartheta,p):=E(y^i(\vartheta,p),\vartheta),\quad \overline{E}^i(\vartheta):=\widetilde{E}^i(\vartheta,0),$$

and noting that $\overline{E}$, as given in (7.3), can be defined via $\overline{E}(\vartheta)=\widetilde{E}(\vartheta,0)$.

Corollary 7.30. *The function $\overline{E}$ is piecewise differentiable and its Clarke subdifferential $\partial^C\overline{E}(\vartheta)$ is given by*

$$\partial^C\overline{E}(\vartheta)=\operatorname{co}\left\{\nabla_x E(x(\vartheta),\vartheta)\,\mathrm{d}_\vartheta y^i(\vartheta,0)+D_\vartheta E(x(\vartheta),\vartheta)\ :\ i\in I(\vartheta)\right\}. \tag{7.14}$$

Proof. By the chain rule for piecewise smooth functions Proposition 7.6, $\widetilde{E}$ is piecewise differentiable, and therefore also locally Lipschitz continuous. By Proposition 2.32 *(iii)*, the Clarke subdifferential corresponds to the convex hull of the set

$$S=\left\{(v,w)\in\mathbb{R}^{m,n}\ \middle|\ \begin{array}{l}(v,w)=\lim_{k\to\infty}D\overline{E}(\vartheta^k,p^k),\ \text{such that } (\vartheta^k,p^k)\to(\vartheta,p)\\ \text{and } \widetilde{E} \text{ is differentiable at } (\vartheta^k,p^k)\end{array}\right\}$$

Since $\widetilde{E}$ is piecewise differentiable with minimal local representation of $\widetilde{E}^i$, $i\in I(\vartheta)$, we derive

$$S=\{D\overline{E}^i(\vartheta^k,p^k)\ :\ i\in I(\vartheta)\}.$$

Finally, by [54, Theorem 2.3.10]

$$\partial^C\overline{E}(\vartheta)=\partial^C\widetilde{E}(\vartheta,0)\circ[I_m,0_n]^T=\operatorname{co}\{D\overline{E}^i(\vartheta)\ :\ i\in I(\vartheta)\}.$$

The expression in (7.14) is obtained by applying the chain rule to E^i. □

7.4.3 Examples of bilevel problems

We discuss examples of lower-level objective functions F that satisfy the criteria in Assumption 7.23. First we give examples.

Partly smooth functions

As the issue of strong convexity is separate from the other assumptions, we assume for simplicity that F is strongly convex, and focus on examples that satisfy the remaining criteria.

We primarly consider variational regularisation problems of the form

$$F(x,\vartheta) := V(x,\vartheta) + R(x,\vartheta),$$

where $V : \mathbb{R}^n \times \Omega \to \overline{\mathbb{R}}$ represents the data fidelity term and $R : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ the regularisation term, as described in Section 1.1.2.

Furthermore, note that if F satisfies Assumption 7.23 and $G : \mathbb{R}^n \times \Omega \to \mathbb{R}$ is C^2-smooth such that $G(\cdot,\vartheta)$ is convex for each ϑ, then $F+G$ also satisfies Assumption 7.23.

We first consider the important class of *polyhedral* functions, which include the ℓ^1 norm, the ℓ^∞ norm, and any linear precomposition of these norms, e.g. $\|K\cdot\|_1$ for $K \in \mathbb{R}^{l,n}$ (thus including anistropic total variation). A polyhedral function $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ is any function which can be written in the form

$$R(x) = \max_{i=1,\dots,N} \{\langle a^i, x\rangle - b_i\} + \delta_{\cap_{i=N+1}^{M}\{x : \langle a^i,x\rangle - b_i\}}(x).$$

As is laid out in [130, Example 3.4], such functions are partly smooth relative to linear manifolds, and the partition of manifolds clearly satisfy Assumption 7.23 *(A3-A4)*. Thus, if $R : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ is such that $R(\cdot,\vartheta)$ is polyhedral, so that the linear manifolds $\mathcal{M} \in \mathfrak{M}_x$ are invariant with respect to the parameters ϑ, then R satisfies the regularity assumptions in Assumption 7.23. Since polyhedral functions are piecewise linear, their Riemannian Hessian vanishes, i.e. $\nabla^2_{\mathcal{M}} R = 0$.

Another important class of functions are *general group Lasso* functions, which are of the form

$$R(x) := \sum_{i=1}^{N} \|B_i x\|,$$

where $\{B_i \in \mathbb{R}^{l,n}\}_{i=1}^N$ is a collection of matrices. To show that these functions are partly smooth relative to a collection of linear manifolds, we proceed accordingly. The function $\|B\cdot\|$ is smooth on $\mathbb{R}^n \setminus \ker B$ and partly smooth at $x \in \ker B$ relative to $\ker B$. By [215, Proposition 9], sums of partly smooth functions with respect to linear manifolds are also partly smooth, and this includes the group Lasso function. Provided that $\ker B_i$ is invariant with respect to the parameters ϑ, the group Lasso satisfies the regularity properties in Assumption 7.23. Note that group Lasso functions include the isotropic total variation seminorm.

For the Riemannian Hessian of general group Lasso functions, we consider a simpler group Lasso function where B_i represents an orthogonal projection onto the subspace $X_i \subset \mathbb{R}^n$. In this case, R is partly smooth at x relative to $\mathcal{M} = \mathbb{R}^n \setminus (\cup_{x \in X_i^\perp} X_i)$, and for $v \in \mathcal{M}$,

$$\nabla^2_{\mathcal{M}} R(x)v = \sum_{i\,:\, X_i \subset \mathcal{M}} \frac{1}{\|B_i x\|} B_i v - \frac{\langle B_i x, B_i v\rangle}{\|B_i x\|^2} B_i x.$$

Next, we give an example of a function which is partly smooth relative to a nonlinear manifold, namely the nuclear norm,

$$\|\cdot\|_* : x \in \mathbb{R}^{n_1,n_2} \mapsto \|\sigma(x)\|_1,$$

where $\sigma(x)$ is the vector of singular values of the matrix x. The nuclear norm can be viewed as the convex relaxation of the rank of a matrix [86], and it is partly smooth [62, 135] relative to the constant r-rank manifold [132]

$$\{x \in \mathbb{R}^{n_1,n_2} \ : \ \operatorname{rank}(x) = r\}.$$

Thus one might consider

$$R(x,\vartheta) = \vartheta \|x\|_*,$$

for $\vartheta \geq 0$. For its Riemannian Hessian, gradient, and Weingarten map, see [214, Example 21].

Finally, indicator functions δ_C for $C \subset \mathbb{R}^n$ are partly smooth if C is a polytope, i.e. if δ_C is polyhedral, or if C has a smooth boundary.

Applications to bilevel problems

We proceed to discuss some bilevel problems which fit within this framework.

A natural example is that of weighting of regularisation terms,

$$F : \mathbb{R}^n \times (0,\infty)^m \to \overline{\mathbb{R}}, \quad F(x,\vartheta) = V(x,\vartheta) + \sum_{i=1}^{m} \vartheta_i R_i(x),$$

where V is C^2-smooth and $\sum_{i=1}^{m} R_i$ satisfies *(A2)*.

Another example is that of learning sampling patterns for compressed sensing in MRI [205], for which we can consider the model

$$F(x,\vartheta) := \frac{1}{2}\|S(\vartheta)(Kx - f^{\delta})\|^2 + \alpha(\vartheta)R(x) + \delta_{\geq 0}(x) + \frac{\varepsilon}{2}\|x\|^2, \tag{7.15}$$

where $\alpha(\vartheta) = \vartheta_m$ and $\vartheta \mapsto S(\vartheta)$ is a matrix-valued function given by

$$S(\vartheta) = \operatorname{diag}((\vartheta)_{i=1}^{m-1}),$$

and $\Omega = [0,1]^{m-1} \times [0,\infty)$. Here $\operatorname{diag}(\vartheta)$ denotes the diagonal matrix with diagonal values given by the vector ϑ. Furthermore $\varepsilon > 0$ and the corresponding term is included to enforce strong convexity. The objective function F clearly satisfies Assumption 7.23 if R is given by any of the examples mentioned above. In [205] they primarily consider the total variation seminorm.

Next we consider two forms for dictionary learning problems, as discussed in [177]. The lower-level objective function for synthesis-based priors is given by

$$x_{\vartheta} \in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \frac{1}{2}\|K(\vartheta)x - f^{\delta}\|^2 + R(x,\vartheta),$$

where $\vartheta \mapsto K(\vartheta) \in \mathbb{R}^{l,n}$ is a C^1-smooth matrix-valued function, and R is a regularisation term, for example $R(\cdot,\vartheta) = \vartheta_m\|\cdot\|_1$. While F varies smoothly with the parameters, in practice $K(\vartheta)$ could have a nontrivial kernel, so $F(\cdot,\vartheta)$ would not necessarily be strongly convex. In these cases, one might want to consider weakening the strong convexity assumption.

On the other hand, the lower-level objective function for analysis-based priors is given by

$$x_{\vartheta} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \frac{1}{2}\|x - f^{\delta}\|^2 + R(K^*(\vartheta)x,\vartheta). \tag{7.16}$$

In this case, the data fidelity term enforces strong convexity of $F(\cdot,\vartheta)$. However, F is no longer differentiable with respect to the parameters ϑ, for example if

$$R(K^*(\vartheta)x,\vartheta) = \|K^*(\vartheta)x\|_1,$$

and letting ϑ freely parametrise the elements of K. Thus the partial smoothness framework for bilevel optimisation is not directly applicable to (7.16). However, we propose to work around this by reformulating (7.16) as the saddle point problem

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^l}\frac{1}{2}\|x-f^\delta\|^2+\langle x,K(\vartheta)y\rangle-R^*(y,\vartheta), \tag{7.17}$$

where $R^*(\cdot,\vartheta)$ is the convex conjugate of $R(\cdot,\vartheta)$ with respect to the first variable. By separating the linear operator from the regularisation term, we obtain a problem for which the parameters ϑ vary smoothly. It remains to extend the concepts of implicit differentiation under partial smoothness to saddle point problems, and consider algorithmic differentiation of primal-dual optimisation methods, something which we discuss in the outlook Section 8.3.3.

7.5 Algorithmic differentiation

In this section, we consider algorithmic differentiation of forward–backward splitting algorithms, in order to differentiate $x(\cdot)$. As mentioned in the literature, previous works including [166, 167, 68] have studied algorithmic differentiation for nonsmooth variational problems in image processing, however these did not look at convergence guarantees for the derivatives. In this section, we seek to establish conditions under which the derivatives converge, and when they might fail to do so.

We therefore suppose that the lower-level objective function F can be split as

$$F(x,\vartheta)=V(x,\vartheta)+R(x,\vartheta), \tag{7.18}$$

where $V,R\in\Gamma_0(\mathbb{R}^n)$ and V is L-smooth.

We consider the following setting. For a parameter choice ϑ and starting point $x^0\in\mathbb{R}^n$, the iterates of the algorithm are given by

$$x^{k+1}(\vartheta):=\mathcal{A}(x^k(\vartheta),\vartheta),\quad k\in\mathbb{N}. \tag{7.19}$$

Hence we view each iterate x^k as a function of ϑ, where $x^0(\vartheta)\equiv x^0$ is constant.

We assume that $\mathcal{A}$ is piecewise C^1-smooth, which we show in the next section holds provided the lower-level objective function satisfies Assumption 7.23. In this case, one can use the chain rule Proposition 7.6 and differentiate (7.19) to recursively compute the derivative

$$Dx^{k+1}(\vartheta)=\nabla_x\mathcal{A}(x^k(\vartheta),\vartheta)Dx^k(\vartheta)+D_\vartheta\mathcal{A}(x^k(\vartheta),\vartheta). \tag{7.20}$$

7.5.1 Proximal maps

Proximal maps are key to nonsmooth optimisation methods. For the algorithms we consider, the nonsmoothness of the iterative map $\mathcal{A}$ is induced via proximal maps acting on the nonsmooth term $R(\cdot,\vartheta)$. In what follows, we therefore consider separately the differentiation of these maps.

For $R \in \Gamma_0(\mathbb{R}^n)$ and $\tau > 0$, we define the Moreau function accordingly,

$$R_\tau(x) := \min_{y\in\mathbb{R}^n} \tau R(y) + \frac{1}{2}\|y-x\|^2. \tag{7.21}$$

The corresponding proximal map is defined as

$$\operatorname{prox}_{\tau R}(x) := \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \tau R(y) + \frac{1}{2}\|y-x\|^2. \tag{7.22}$$

Now suppose the lower-level objective function is given by (7.18) and satisfies Assumption 7.23, and consider the parameter-dependent proximal map

$$\operatorname{prox}_{\tau R}(y,\vartheta) := \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \tau R(x,\vartheta) + \frac{1}{2}\|x-y\|^2. \tag{7.23}$$

Since V is smooth and $V+R$ satisfies Assumption 7.23, the function

$$f_{\tau R}(x,\vartheta,y) := \frac{1}{2}\|x-y\|^2 + \tau R(x,\vartheta), \tag{7.24}$$

also satisfies these assumptions, treating (ϑ, y) as the parameters. Denote by $I_{\tau R}(\vartheta,y)$ the corresponding index set for this (7.24), as defined in (7.13).

The following result is immediate when one observes that $\operatorname{prox}_{\tau R}(y,\vartheta)$ is simply the solution map corresponding to (7.24).

Lemma 7.31. *Suppose $F : \mathbb{R}^n \times \Omega \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is given by* (7.18) *and satisfies Assumption 7.23. Then the parameter dependent proximal map given by* (7.23) *is piecewise smooth in both arguments, with differential $D\operatorname{prox}_{\tau R}(y,\vartheta)$ having a minimal local representation of*

$$\left\{ \begin{bmatrix} (P_{T_x\mathcal{M}_i} + \tau\nabla^2_{\mathcal{M}_i} R(x,\vartheta))^\dagger \\ -\tau(P_{T_x\mathcal{M}_i} + \tau\nabla^2_{\mathcal{M}_i} R(x,\vartheta))^\dagger D_\vartheta \nabla_{\mathcal{M}} R(x,\vartheta) \end{bmatrix} \;\middle|\; i \in I_{\tau R}(\vartheta,y) \right\}, \tag{7.25}$$

where $x = prox_{\tau R}(y,\vartheta)$.

Proof. We apply Theorem 7.29 to (7.24) which gives us the result. □

Lemma 7.31 will be central to the framework of algorithmic differentiation of forward-backward splitting methods. The following result relates nondegeneracy of the variational problem (7.1) to local C^1-smoothness of $\operatorname{prox}_{\tau R}$ around $(\vartheta, x(\vartheta))$.

Proposition 7.32. *Consider the setting of Lemma 7.31 and suppose furthermore that* (ND) *holds for* $F(\cdot, \vartheta^*)$ *at* x^*. *Then* $\operatorname{prox}_{\tau R}(\vartheta, y)$ *is continuously differentiable near* $(\vartheta^*, x^* - \tau\nabla_x V(x^*, \vartheta^*))$.

Proof. It is sufficent to show that $0 \in \operatorname{ri} \partial_x f_{\tau R}(x^*, \vartheta, x^* - \tau\nabla_x V(x^*, \vartheta^*))$ and apply [130, Theorem 5.7]. Writing out the subdifferential of $f_{\tau R}$ gives us

$$\partial_x f_{\tau R}(x^*, \vartheta^*, x^* - \tau\nabla_x V(x^*, \vartheta^*)) = \tau(\nabla_x V(x^*, \vartheta^*) + \partial_x R(x^*, \vartheta^*)) = \tau\partial_x F(x^*, \vartheta^*),$$

and the result follows. □

7.5.2 Forward-backward-type methods

We are now ready to introduce our class of forward-backward algorithms and study the corresponding algorithmic derivatives.

Let the lower-level objective function be given by (7.18), i.e. $F(x, \vartheta) = V(x, \vartheta) + R(x, \vartheta)$ where V is L-smooth with respect to x.

Algorithm 5 Forward-backward splitting method

Input: starting point $x^0 = x^{-1} \in \mathbb{R}^n$, parameter $\vartheta \in \Omega$, time steps $(\tau_k)_{k\in\mathbb{N}} \subset [\overline{\varepsilon}, 2/L - \overline{\varepsilon}]$ for some $\overline{\varepsilon} > 0$, inertial parameters $(a_k)_{k\in\mathbb{N}} \subset [0, 1]$,

```
for $k = 0, 1, 2, \dots$ do
    $y^k = x^k + a_k(x^k - x^{k-1})$
    $x^{k+1} = \operatorname{prox}_{\tau_k R}(y^k - \tau_k\nabla_x V(y^k, \vartheta), \vartheta)$        (7.26)
end for
```

As pointed out in [135], this class of algorithms covers the original forward-backward method [138] when $a_k = 0$, and variants of accelerated FISTA (fast iterative shrinkage-thresholding algorithm) [8, 14, 49] when $\tau_k \in [\overline{\varepsilon}, 1/L]$ and $a_k \to 1$. Note that Algorithm 5 is slightly more restrictive than the class of forward-backward algorithms considered in [135], i.e. we fix $a_k = b_k$ in their Algorithm 1.

Recall from (7.19) that we represent the update in Algorithm 5 by a mapping $\mathcal{A} : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$. However, since our class of algorithms includes multistep mappings, as well as time steps and inertial parameters that depend on the iteration index k, we adapt the notation in (7.19) accordingly. We denote by $\mathcal{A}_k : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ and require that $\tau_k \to \tau$, $a_k \to a$, so that $\mathcal{A}_k \to \mathcal{A}$. Additionally, we rewrite the multistep iterative procedure as a single step accordingly. The algorithm update (7.26) is equivalent to

$$\begin{gathered} z^{k+1}(\vartheta) = \mathcal{A}_k(z^k(\vartheta), \vartheta), \\ z^k := \begin{bmatrix} x^k \\ x^{k-1} \end{bmatrix}, \quad \mathcal{A}_k(z, \vartheta) := \begin{bmatrix} f_k\left(g_k(h_k(z^k), \vartheta), \vartheta\right) \\ z_1^k \end{bmatrix}, \end{gathered} \tag{7.27}$$

where $f_k(x, \vartheta) = \operatorname{prox}_{\tau_k R}(x, \vartheta)$, $g_k(x, \vartheta) = x - \tau_k \nabla_x V(x, \vartheta)$, and $h_k([z_1, z_2]^T) = z_1 + a_k(z_1 - z_2)$. Clearly g_k and h_k are differentiable, so by Lemma 7.31 and the chain rule, the algorithmic mapping $\mathcal{A}$ in (7.27) is piecewise differentiable.

We now proceed to the main result of this section.

Theorem 7.33. *Let the function $F \equiv V + R : \mathbb{R}^n \times \Omega \to \overline{\mathbb{R}}$ be given by* (7.18) *and suppose it satisfies Assumption 7.23. Furthermore, suppose for $\vartheta \in \Omega$ that the iterates $x^k(\vartheta)$ given by* (7.26) *converge to a minimiser x^* of $F(\cdot, \vartheta)$, and that* (ND) *holds for $F(\cdot, \vartheta)$ at x^*. Then the sequence of (semi)derivatives $Dx^k(\vartheta)$ converges linearly to the single-valued limit $Dx(\vartheta)$.*

Remark 7.34. *Of course, one can also differentiate with respect to algorithmic parameters such as τ_k and a_k. We have not considered this, as our motivation is to study the convergence of the algorithmic derivatives to the implicit derivative* (7.4)*, the latter of which does not involve algorithmic parameters.*

We also point out that convergence of the iterates x^k remains an open problem for some variants of FISTA, we refer to the discussion in [135]. Of course, as we assume strong convexity, we can also deduce that $F(x^k) \to F(x(\vartheta)) \implies x^k \to x(\vartheta)$.

Proof. There is a smooth manifold $\mathcal{M}$ such that F is partly smooth at x^* relative to $\mathcal{M}$. Since $x^k(\vartheta) \to x^*$, we have the limit

$$g_k(h_k(z^k), \vartheta) \to x^* - \tau \nabla_x V(x^*, \vartheta).$$

Since x^* is a strong critical point of $F(\cdot, \vartheta)$, Proposition 7.32 implies that there exists $K \in \mathbb{N}$ such that for all $k \geq K$, f_k is locally continuously differentiable around $g_k(h_k(z^k), \vartheta)$.

Applying (7.20) to (7.27), we have

$$\begin{aligned}
Dz^{k+1}(\vartheta) &= M_k Dz^k(\vartheta) + b^k, \quad \text{where} \\
M_k &:= \begin{bmatrix} (1+a_k)A_k & -a_k A_k \\ I & 0 \end{bmatrix}, \quad A_k := \nabla_x f_k\left(g_k(h_k(z^k),\vartheta),\vartheta\right)\nabla_x g_k(h_k(z^k),\vartheta), \\
b^k &:= \begin{bmatrix} D_\vartheta f_k\left(g_k(h_k(z^k),\vartheta)\right) + \nabla_x f_k\left(g_k(h_k(z^k),\vartheta),\vartheta\right) D_\vartheta g_k(h_k(z^k),\vartheta) \\ 0 \end{bmatrix}.
\end{aligned} \tag{7.28}$$

Denote by f, g, and h the limits of the function sequences f_k, g_k, and h_k. By eventual local differentiability, we have

$$\begin{aligned}
A_k &\to \nabla_x f\left(g(h(z^*),\vartheta),\vartheta\right)\nabla_x g(h(z^*))\nabla h_k(z^*) =: A \in \mathbb{R}^{n,n}, \\
b^k &\to \begin{bmatrix} D_\vartheta f\left(g(h(z^*),\vartheta)\right) + \nabla_x f\left(g(h(z^*),\vartheta),\vartheta\right) D_\vartheta g(h(z^*),\vartheta) \\ 0 \end{bmatrix} =: b \in \mathbb{R}^{n,m}, \\
M_k &\to \begin{bmatrix} (1+a)A & -aA \\ I & 0 \end{bmatrix},
\end{aligned}$$

where $z^* = [x^*, x^*]^T$. By Lemma 7.31, A is given by

$$A = (P_{T_{x^*}\mathcal{M}} + \tau\nabla^2_{\mathcal{M}} R(x^*,\vartheta))^\dagger (I - \tau\nabla^2 V(x^*,\vartheta)).$$

In order to apply Proposition 2.7, we want to show that the spectral radius of M, $\rho(M)$, is less than 1. For this, we first need to show that $\|A\| < 1$. We have

$$\|A\| \le \|(P_{T_{x^*}\mathcal{M}} + \tau\nabla^2_{\mathcal{M}} R(x^*,\vartheta))^\dagger\| \|I - \tau\nabla^2(x^*,\vartheta)\|.$$

By Assumption 7.23, V is μ-convex and R is ν-convex, where either $\mu > 0$ or $\nu > 0$. Since the second matrix above is self-adjoint, by [124, Theorem 9.2.2],

$$\|I - \tau\nabla^2(x^*,\vartheta)\| = \sup_{\|x\|=1} |\|x\|^2 - \tau\langle x, \nabla^2(x^*,\vartheta)x\rangle| \le \max\{|1-\tau\mu|, |1-\bar{\varepsilon}L|\}.$$

For the first matrix,

$$\sup_{\|x\|=1} \|(P_{T_{x^*}\mathcal{M}} + \tau\nabla^2_{\mathcal{M}} R(x^*,\vartheta))^\dagger x\| = \sup_{\|x\|=1,\, x\in T_{x^*}\mathcal{M}} \|(P_{T_{x^*}\mathcal{M}} + \tau\nabla^2_{\mathcal{M}} R(x^*,\vartheta))^\dagger x\|$$

Let $x \in T_{x^*}\mathcal{M}$ and $y = (P_{T_{x^*}\mathcal{M}} + \tau\nabla^2_{\mathcal{M}}R(x^*, \vartheta))^\dagger x$. Then

$$\|x\| = \|y + \tau\nabla^2_{\mathcal{M}}R(x^*, \vartheta)y\| \geq (1+\tau\nu)\|y\|,$$

from which it follows that $\|(P_{T_{x^*}\mathcal{M}} + \tau\nabla^2_{\mathcal{M}}R(x^*, \vartheta))^\dagger\| < 1/(1+\tau\nu)$. Since either μ or ν is strictly positive, it follows that $\|A\| < 1$.

Now suppose $Mz = \lambda z$ for some $z = [z_1, z_2]^T$ and $\lambda \in \mathbb{C} \setminus \{0\}$. We then get

$$\begin{bmatrix} (1+a)A & -aA \\ I & 0 \end{bmatrix} z = \begin{bmatrix} (1+a)z^1 - aAz^2 \\ z^1 \end{bmatrix} = \lambda \begin{bmatrix} z^1 \\ z^2 \end{bmatrix}.$$

Thus $\lambda z^2 = z^1$, which implies that $\lambda z^1 = (1+a-a/\eta)Az^1$, so $\eta/(1+a-a/\eta)$ is a nonzero eigenvalue of A. One can verify that nonzero eigenvalues of A on $\mathbb{C}^n$ coincide with nonzero eigenvalues of A on $T_{x^*}\mathcal{M}$. Furthermore, restricted to this subspace, A satisfies the assumptions of Proposition 2.8, so $\sigma(A) \subset \mathbb{R}$.

If we write $\rho = \|A\| < 1$, then we have $|\lambda/(1+a-a/\lambda)| \leq \rho$. We will show that $|\lambda| < 1$ by case-by-case analysis.

If $\lambda < 0$, then we have $-\lambda < (1+a-a/\lambda)\rho$. One can check that all negative solutions for λ to this lie in $(-1, 0)$. Otherwise, we assume $\lambda > 0$. If $1+a-a/\lambda < 0$, then $\lambda < a/(1+a) < 1$. Otherwise, if $1+a-a/\lambda > 0$, then we have $\lambda^2 - (1+a)\rho\lambda + a\rho < 0$, for which we can show that $\lambda \leq \rho < 1$. Thus $\rho(M) < 1$.

Therefore, by Proposition 2.7, the sequence of derivatives $Dz^k(\vartheta)$ converges linearly to

$$\lim_{k\to\infty} Dz^k(\vartheta) = (I-M)^{-1}b.$$

It remains to show that $Dx(\vartheta)$ solves

$$(I-M)\begin{bmatrix} Dx(\vartheta) \\ Dx(\vartheta) \end{bmatrix} = b.$$

We have

$$(I-M)\begin{bmatrix} Dx(\vartheta) \\ Dx(\vartheta) \end{bmatrix} = \begin{bmatrix} I-(1+a)A & aA \\ -I & I \end{bmatrix}\begin{bmatrix} Dx(\vartheta) \\ Dx(\vartheta) \end{bmatrix} = \begin{bmatrix} (I-A)Dx(\vartheta) \\ 0 \end{bmatrix}.$$

For brevity, write

$$M_V = \nabla^2_{\mathcal{M}} V(x(\vartheta),\vartheta), \quad M_R = \nabla^2_{\mathcal{M}} R(x(\vartheta),\vartheta), \quad D_V = D_\vartheta \nabla_{\mathcal{M}} V(x(\vartheta),\vartheta),$$
$$D_R = D_\vartheta \nabla_{\mathcal{M}} R(x(\vartheta),\vartheta), \quad P = P_{T_{x^*}\mathcal{M}}.$$

Then we have by Lemma 7.26 and Lemma 7.31,

$$\begin{aligned}(I-A)Dx(\vartheta) &= -\Big(I - (P+\tau M_R)^\dagger (I-\tau M_V)\Big)(M_V+M_R)^\dagger (D_V+D_R)\\ &= -\tau(P+\tau M_R)^\dagger(D_V+D_R) = -\tau(P+\tau M_R)^\dagger D_R - \tau(P+\tau M_R)^\dagger D_V = b.\end{aligned}$$

This concludes the proof. □

7.5.3 Bregman proximal methods

The generalisation from the Euclidean distance to Bregman distances is significant to optimisation and regularisation theory. In what follows, we briefly consider the *Bregman proximal method* and show that the derivative convergence result Theorem 7.33 extends to this case under certain conditions.

Denote by $J : \mathbb{R}^n \to \overline{\mathbb{R}}$ a function that is C^1-smooth on $\operatorname{int}\operatorname{dom} J$, and 1-convex.[3] For $F = V + R$ given by (7.18), the iteration map for the Bregman proximal method is given by

$$\begin{aligned}\mathcal{A}_k(y,\vartheta) &:= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} f_k^J(x,\vartheta,y),\\ f_k^J(x,y,\vartheta) &:= \frac{1}{\tau_k} D_J(x,y) + R(x,\vartheta) + \langle \nabla_x V(y,\vartheta), x-y\rangle.\end{aligned} \tag{7.29}$$

Algorithm 6 Bregman proximal method

Input: starting point $x^0 \in \mathbb{R}^n$, parameter $\vartheta \in \Omega$, time steps $(\tau_k)_{k\in\mathbb{N}} \subset [\overline{\varepsilon}, 1/L]$ for some $\overline{\varepsilon} > 0$

```
for $k = 0,1,2,\dots$ do
    $x^{k+1} = \mathcal{A}_k(x^k,\vartheta), \quad \mathcal{A}_k$ given in (7.29)
end for
```

As before, we assume that $\tau_k \to \tau$. In comparison to Algorithm 5, this algorithm is more restrictive, as there is no inertial step, i.e. $a_k = 0$ and $\tau_k \le 1/L$. Regarding the restriction on

[3] We only consider smooth functions, since otherwise the Bregman proximal map would depend on a subgradient choice $p \in \partial J(y)$, further complicating the algorithmic differentiation.

a_k, as is pointed out in [212], FISTA does not seem to be directly extendible to the Bregman distance setting, and while other acceleration variants have been proposed [213], we do not consider these here. Depending on the choice of J, time steps τ_k up to $2/L-\varepsilon$ are possible depending on the Bregman distance generating function J—see [212, Definition 4.1] and surrounding discussion.

Suppose the objective function $V+R$ satisfies Assumption 7.23. Arguing as in Section 7.5.1 and noting the L-smoothness of V, f_k^J satisfies Assumption 7.23, treating (y,ϑ) as the parameters. Denote by $I^J_{\tau_k V,\tau_k R}(y,\vartheta)$ the index set (7.13) corresponding to (7.29). The following result is analogous to Lemma 7.31 and Proposition 7.32 for the Bregman proximal method.

Lemma 7.35. *Suppose* $F=V+R:\mathbb{R}^n\times\Omega\to\overline{\mathbb{R}}$ *is given by* (7.18) *and satisfies Assumption 7.23. Then the Bregman proximal mapping* $\mathcal{A}_k(y,\vartheta)$ *in* (7.29) *is piecewise smooth in both arguments, with differential* $D\mathcal{A}_k(y,\vartheta)=[\nabla_x\mathcal{A}_k(y,\vartheta),D_\vartheta\mathcal{A}_k(y,\vartheta)]^T$ *having a minimal local representation of*

$$\left\{\begin{bmatrix}(\nabla^2_{\mathcal{M}_i}J(x)+\tau_k\nabla^2_{\mathcal{M}_i}R(x,\vartheta))^\dagger(\nabla^2J(y)-\tau_k\nabla^2_xV(y,\vartheta))\\ -\tau_k(\nabla^2_{\mathcal{M}_i}J(x)+\tau_k\nabla^2_{\mathcal{M}_i}R(x,\vartheta))^\dagger(D_\vartheta\nabla_{\mathcal{M}_i}R(x,\vartheta)+D_\vartheta\nabla_xV(y,\vartheta))\end{bmatrix}\right\}_{i\in I_k^J(y,\vartheta)}, \tag{7.30}$$

where $x=\mathcal{A}_k(y,\vartheta)$.

Furthermore, if (ND) *holds for* $F(\cdot,\vartheta)$ *at* x^*, *then* $\mathcal{A}_k$ *is locally continuously differentiable near* (x^*,ϑ).

Proof. Piecewise smoothness follows from Theorem 7.29 applied to $f_k^J(x,y,\vartheta)$.

For the second part, it is sufficent to show that $0\in\operatorname{ri}\partial_x f_k^J(x^*,\vartheta,x^*)$ and apply [130, Theorem 5.7]. We have

$$\partial_x f_k^J(x^*,\vartheta,x^*)=\frac{1}{\tau_k}(\nabla J(x^*)-\nabla J(x^*))+\partial_x R(x*,\vartheta)+\nabla_x V(x^*,\vartheta)=\partial_x F(x*,\vartheta),$$

and the proof is complete. □

Theorem 7.36. *Let the function* $F\equiv V+R:\mathbb{R}^n\times\Omega\to\overline{\mathbb{R}}$ *be given by* (7.18) *and suppose it satisfies Assumption 7.23. Furthermore, suppose for* $\vartheta\in\Omega$ *that the iterates* $x^k(\vartheta)$ *given by Algorithm 6 converges to a minimiser* $x^*\in\operatorname{int}\operatorname{dom}J$ *of* $F(\cdot,\vartheta)$, *and that* (ND) *holds for* $F(\cdot,\vartheta)$ *at* x^*. *Then the sequence of (semi)derivatives* $Dx^k(\vartheta)$ *converges linearly to the single-valued limit* $Dx(\vartheta)$.

Proof. We argue along the same lines as in the proof to Theorem 7.33. Let $\mathcal{M} \subset \mathbb{R}^n$ be a smooth manifold such that F is partly smooth at (x^*, ϑ) relative to $\mathcal{M} \times \mathbb{R}^n$. By Lemma 7.35, there is $K \in \mathbb{N}$ such that for all $k \geq K$, f_k is continuously differentiable near $g_k(x^k, \vartheta)$.

Applying (7.20) to (7.29), we have

$$Dx^{k+1}(\vartheta) = A_k Dx^k(\vartheta) + b^k, \tag{7.31}$$

where

$$A_k := \nabla_x f_k^J(x^k, \vartheta), \quad b^k := D_\vartheta f_k^J(x^k, \vartheta)$$

Write $f^J := \lim_{k\to\infty} f_k^J$. By Lemma 7.35, there is $K \in \mathbb{N}$ such that for all $k \geq K$, the iterations $\mathcal{A}_k(x^k, \vartheta)$ are locally continuously differentiable, and we have

$$\begin{aligned} A_k &\to (\nabla^2_{\mathcal{M}} J(x^*) + \tau \nabla^2_{\mathcal{M}} R(x^*, \vartheta))^\dagger (\nabla^2 J(x^*) - \tau \nabla^2_x V(x^*, \vartheta)) =: A \in \mathbb{R}^{n,n}, \\ b^k &\to -\tau(\nabla^2_{\mathcal{M}} J(x^*) + \tau \nabla^2_{\mathcal{M}} R(x^*, \vartheta))^\dagger (D_\vartheta \nabla_{\mathcal{M}} R(x^*, \vartheta) + D_\vartheta \nabla_x V(x^*, \vartheta)) =: b \in \mathbb{R}^{n,m}. \end{aligned}$$

Write for shorthand

$$M_J := \nabla^2_{\mathcal{M}} J(x^*), \quad M_R := \nabla^2_{\mathcal{M}} R(x^*, \vartheta), \quad M_V := \nabla^2_x V(x^*, \vartheta),$$

so that $A = (M_J + \tau M_R)^\dagger (M_J - \tau M_V)$.

We need to show that $\rho(A) < 1$. Suppose $Ax = \lambda x$ for some $x \in \mathbb{C}^n$, $\lambda \in \mathbb{C} \setminus 0$. Note that any eigenvector x of A must lie in the subspace $T_{x^*}\mathcal{M}$, so the spectrum of A in $\mathbb{R}^n$ coincides with its spectrum restricted to $T_{x^*}\mathcal{M}$. Furthermore, restricted to this subspace, A satisfies the conditions for Proposition 2.8, meaning $\lambda \in \mathbb{R}$.

Since $x \in T_{x^*}\mathcal{M}$, we can rearrange $\lambda x = Ax$ to get

$$(1-\lambda) M_J x = \tau(\lambda M_R - M_V)x$$

Taking the inner product on each side with respect to x, we get

$$(1-\lambda)\langle x, M_J x\rangle = \tau\lambda\langle x, M_R x\rangle + \tau\langle x, M_V x\rangle. \tag{7.32}$$

By strong convexity of F and J, there is $\mu, \nu \geq 0$ with $\mu + \nu > 0$ such that $\langle x, M_J x\rangle \geq \|x\|^2$, $\tau\langle x, M_R x\rangle \geq \tau\nu\|x\|^2$, and $\tau\langle x, M_V x\rangle \in [\bar{\varepsilon}\mu\|x\|^2, \|x\|^2]$. One can then verify that for (7.32) to hold, $\lambda \in [0, 1)$.

Therefore, by Lemma 7.35, $Dx^k(\vartheta)$ converges linearly to $(I-A)^{-1}b$. It remains to show that $(I-A)Dx(\vartheta)=b$. Writing

$$D_V = D_\vartheta \nabla_{\mathcal{M}} V(x(\vartheta),\vartheta), \quad D_R = D_\vartheta \nabla_{\mathcal{M}} R(x(\vartheta),\vartheta),$$

we have

$$\begin{aligned}(I-A)Dx(\vartheta) &= -\Big(I-(M_J+\tau M_R)^\dagger(M_J-\tau M_V)\Big)(M_V+M_R)^\dagger(D_V+D_R)\\ &= -\tau(M_J+\tau M_R)^\dagger(D_V+D_R)=b.\end{aligned}$$

This concludes the proof. □

As mentioned earlier, we do not consider nonsmooth Bregman distance generating functions $J:\mathbb{R}^n\to\overline{\mathbb{R}}$, as this would involve differentiation with respect to an additional variable, namely subgradients $p^k\in\partial J(x^k)$. We therefore leave this for future research.

Second, in Theorem 7.36, we assume that $x^*\in\operatorname{int}\operatorname{dom}J$. This ensures that A_k converges to a unique limit. However, this assumption does not hold in general, including for some popular Bregman distances such as the Kullback–Leibler divergence $D_J(x,y)=x(\log x-\log y)-(x-y)$ generated by the entropy function $J(x)=x\log x$ (in one dimension). Furthermore, as was demonstrated in [166, 167], one can achieve iterative methods that solve nonsmooth variational methods, yet whose iterative map $\mathcal{A}(x,\vartheta)$ is continuously differentiable, provided the nonsmoothness can be expressed as convex constraints that coincide with $\operatorname{cl}\operatorname{dom}J$. In these settings, one expects $x^*\notin\operatorname{dom}J$.

While we do not prove convergence results for the case where $x^*\notin\operatorname{int}\operatorname{dom}J$, we show for a simple example with the Kullback–Leibler divergence that the algorithmic iterates Dx^k do converge to the implicit derivative Dx even when $x^*=0\notin\operatorname{dom}J$.

Example 7.37. *Consider a simple example*

$$x(\vartheta)=\operatorname*{arg\,min}_{x\in\mathbb{R}^n} V(x,\vartheta)+\delta_{\geq 0}(x),$$

and $J(x)=\sum_{i=1}^n x_i\log x_i$. The Bregman distance is the Kullback–Leibler divergence given by

$$D_J(x,y)=\sum_{i=1}^n x(\log x-\log y)-(x-y).$$

We assume that $x^0\in\mathbb{R}^n$ is such that $\{x\ :\ V(x)\leq V(x^0)\}\subset[0,1]^n$, as J this ensures that J is 1-convex for all x^k. In general, one can rescale J to ensure 1-convexity on greater domains.

For $\tau_k \in [\overline{\varepsilon}, 1/L]$, the iterates of Algorithm 6 yield the updates

$$x^{k+1}(\vartheta) = x^k \exp(-\tau \nabla V(x^k(\vartheta), \vartheta)) \to x(\vartheta) =: x^*.$$

We differentiate this with respect to ϑ and obtain

$$Dx^{k+1}(\vartheta) = Dx^k(\vartheta) \exp(-\tau \nabla V(x^k(\vartheta), \vartheta)) - x^k D_\vartheta \left(\exp(-\tau \nabla V(x^k(\vartheta), \vartheta)) \right),$$

where exp *is applied element-wise to the vectors. For each i, if $x_i^k \to 0$, then*

$$[Dx^{k+1}(\vartheta)]_i = Dx^k(\vartheta) \exp(-\tau \nabla_i V(x^k(\vartheta), \vartheta)) + O(\|x^k\|).$$

In this case, the condition (ND) *holds if and only if, for each i such that $x_i^* = 0$, one has $[\nabla V(x^*, \vartheta)]_i > 0$. In this case, we see that $[Dx^k]_i \to 0$ linearly. In conclusion, we have*

$$Dx^k(\vartheta) \to Dx(\vartheta).$$

7.5.4 Failure of convergence under the degenerate case

In the previous section, we proved under the nondegeneracy assumption that the algorithmic derivative converges to the true derivative. Now we consider what might happen when the nondegeneracy condition does not hold. We will show that in this case, the sequence, and any subsequence, of algorithmic derivatives can fail to converge to a subgradient of $x(\cdot)$.

In such cases, the solution $x(\vartheta^*)$ may be at the transition point between two manifolds, across which the Dx behaves discontinuously. In Theorem 7.29, we proved that $x(\cdot)$ is piecewise continuous and therefore semidifferentiable at such points. One could therefore hope that the sequence, or a subsequence, of algorithmic derivatives converges to a Clarke subgradient, or, equivalently by (2.1), to a convex combination of the local gradient representatives Dx_i. This would indeed be the case if the iterates identified and remained within one of the manifolds. However, it is also possible that the iterates oscillate between different manifolds. We first demonstrate with a numerical example, and then justify it mathematically.

We consider a problem $F(x, \vartheta) : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$ given by

$$\frac{1}{2}\|Ax - b\|^2 + \vartheta \|x\|_1,$$

where $A \in \mathbb{R}^{3,3}$ and $b \in \mathbb{R}^3$ are randomly generated, and approximate ϑ to an accuracy of 10^{-14} such that (ND) fails. We then run the standard forward-backward algorithm (with $a^k = 0$) and $\tau = 1.8/L$, compute the algorithmic derivative at each iterate, and measure the

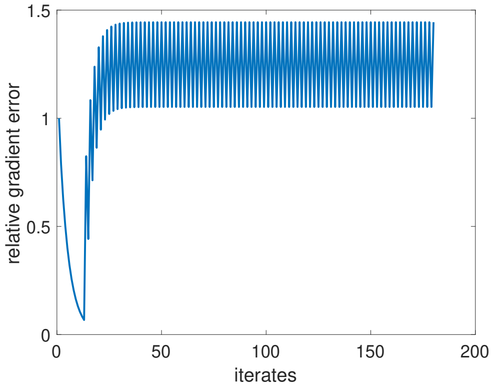


Fig. 7.1 The relative distance of the algorithmic derivative to the convex hull of local derivatives with respect to iterates, i.e. $\operatorname{dist}(Dx^k, \partial x)/\operatorname{dist}(Dx^1, \partial x)$.

distance of the algorithmic derivative to the convex hull of local derivatives $\partial x = \operatorname{co}\{D\widetilde{x}^i : i \in I(\vartheta)\}$. See Figure 7.1 for a plot of $\operatorname{dist}(Dx^k, \partial x)/\operatorname{dist}(Dx^1, \partial x)$ with respect to the iterates. The iterates quickly converge to the minimiser, after which they oscillate between the two manifolds. Similarly, the algorithmic derivatives alternate between two vectors, neither of which lie close to the subdifferential ∂x. The computation was done on MATLAB.

We will now provide a mathematical justification for why the automatic derivatives fail to approximate any (sub)gradient. For this, we consider a simplified scenario, where there are two distinct, linear manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $x \in \mathcal{M}_1 \cap \operatorname{cl} \mathcal{M}_2$, and such that eventually $x^{2k} \in \mathcal{M}_1$ and $x^{2k+1} \in \mathcal{M}_2$ for all $k \geq K$. Furthermore, write

$$\begin{aligned} A_1 &= \lim_{k\to\infty} D_x\mathcal{A}(x^{2k}, \vartheta), & b^1 &= \lim_{k\to\infty} D_\vartheta\mathcal{A}(x^{2k}, \vartheta), \\ A_2 &= \lim_{k\to\infty} D_x\mathcal{A}(x^{2k+1}, \vartheta), & b^2 &= \lim_{k\to\infty} D_\vartheta\mathcal{A}(x^{2k+1}, \vartheta). \end{aligned}$$

Then, provided $\|A_1\| < 1$, $\|A_2\| < 1$, the algorithmic derivatives behave asymptotically as

$$Dx^{k+1} = \begin{cases} A_1 Dx^k + b_1 + o(\rho^k), & \text{if } k \text{ is even,} \\ A_2 Dx^k + b_2 + o(\rho^k), & \text{if } k \text{ is odd,} \end{cases}$$

where $\rho \in (\max\{\|A_1\|, \|A_2\|\}, 1)$. In the limit, these iterates converge to

$$\lim_{k\to\infty} Dx^{2k} = (I - A_2A_1)^{-1}(A_2b_1 + b_2), \qquad \lim_{k\to\infty} Dx^{2k+1} = (I - A_1A_2)^{-1}(A_1b_2 + b_1). \quad (7.33)$$

The two local derivatives are given by

$$D\widetilde{x}^i = -(\nabla^2_{\mathcal{M}_i} F)^\dagger D_\vartheta \nabla_{\mathcal{M}_i} F = (I - A_i)^{-1} b_i, \quad i = 1, 2. \quad (7.34)$$

It remains to note that in general, the limits in (7.33) are not a linear combination of the limits in (7.34).

Suppose F is continuous on $\operatorname{dom} F$. As $\mathcal{M}_1 \subset \operatorname{cl} \mathcal{M}_2$, a C^2-smooth representative of F along $\mathcal{M}_2$, $\widehat{F}$, is therefore also a representative along $\mathcal{M}_1$. Therefore $\nabla^2_{\mathcal{M}_1} F = P \nabla^2_{\mathcal{M}_2} F P$, where P is the projection onto $T_{\mathcal{M}_1}(x)$. Using block matrix notation and changing the basis, we have

$$\nabla^2_{\mathcal{M}_2} F = \begin{bmatrix} B & C & 0 \\ C^T & D & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

for some matrices B and C, and where $D = \pi \nabla^2_{\mathcal{M}_2} F$ is the projection to its intrinsic subspace, meaning D is positive-definite. By Rhode's theorem on pseudoinverses of Hermitian block matrices [193],

$$(\nabla^2_{\mathcal{M}_2} F)^\dagger = \begin{bmatrix} B^\dagger + B^\dagger C Q^\dagger C^T B^\dagger & -B^\dagger C Q^\dagger & 0 \\ -Q^\dagger C^T B^\dagger & Q^\dagger & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $Q = D - C^T B^\dagger C$. We may compare this to the pseudoinverse of $\nabla^2_{\mathcal{M}_2} F$, which is given by

$$(\nabla^2_{\mathcal{M}_1} F)^\dagger = \begin{bmatrix} 0 & 0 & 0 \\ 0 & D^\dagger & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Plugging these expressions into (7.33) and (7.34), it becomes clear that the algorithmic derivatives will in general not converge to a subdifferential of the true derivative.

7.6 Numerical experiments

In what follows, we test the framework for some simple examples. We emphasise that we only consider the computation of $Dx(\vartheta)$, and leave solving the actual bilevel problem to future work. All numerics are done on MATLAB. For all of the examples, we are able to ensure that the minimiser satisfies (ND) and that we can identify the active manifold, get sufficiently close to the minimiser, and compute the implicit derivative $Dx(\vartheta)$ to a high order of accuracy. For the numerical results, we can therefore reliably compare with the 'true derivative'. Note that in practice, this is often not the case. See Section 8.3.3 for a further discussion of this.

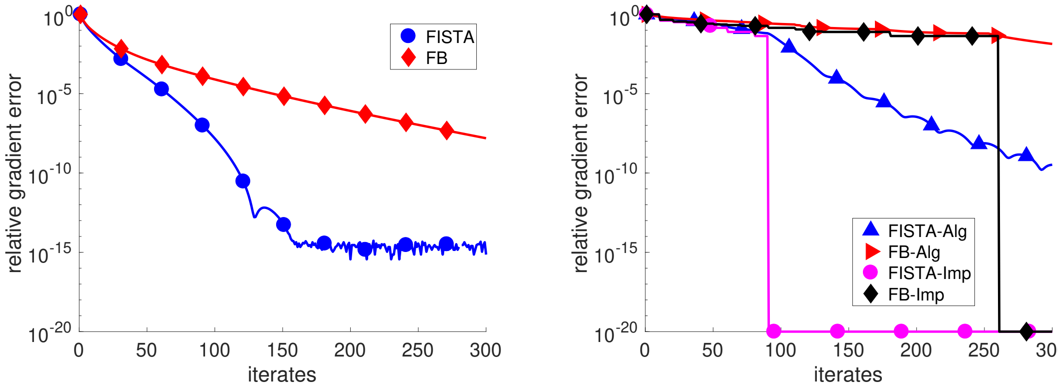


Fig. 7.2 Left: The relative objective of (7.35) with respect to the iterates, for FISTA and FB. Right: The relative error $\|Dx^k(\vartheta)-Dx(\vartheta)\|/\|Dx^1(\vartheta)-Dx(\vartheta)\|$ with respect to the iterate, for FISTA and FB, and implicit (Imp) and algorithmic (Alg) differentiation.

We first compare algorithmic and implicit differentiation of the solution mapping corresponding to the lower-level objective function

$$F(x,\vartheta)=\frac{1}{2}\|Ax-f^{\delta}\|^2+\vartheta\|x\|_1, \tag{7.35}$$

where $\vartheta=10$, and $A\in\mathbb{R}^{800,800}$ and $f^{\delta}\in\mathbb{R}^{800}$ are generated by independent Gaussian draws. We solve using FISTA, with inertial parameter $a_k=(k-1)/(k+30)$, and FB, which corresponds to $a_k=0$, with the time step $\tau_k=1/\|A\|^2$ in both cases. At each iterate, we compute the algorithmic derivative and the implicit derivative at this stage, the latter computed as if x^k were the actual minimiser of (7.35).

See Figure 7.2 for the results. The derivative of $x(\vartheta)$ equals $P(A^*A)^{-1}P$, where P is the projection onto subspace spanned by the basis vectors e^i for which $[x(\vartheta)]_i\neq 0$. Therefore, the implicit derivative remains unchanged for all x with the same support. We therefore observe that the error in the implicit derivative vanishes when the iterate identifies the correct support. In contrast, the algorithmic derivative converges to the true derivative at a linear rate relating to the condition number of A^*A.

We run the same experiment as above, with regularised logistic regression of the form

$$F(x,\vartheta)=\sum_{i=1}^{l}\log\Big(1+\exp(-y_i\langle w^i,x\rangle)\Big)+\vartheta\|x\|_1. \tag{7.36}$$

Here $W\in\mathbb{R}^{n,l}$ is randomly generated from independent Gaussian draws, and $y\in\{\pm1\}^l$ with each element taking either value with equal probability. We set $n=200$, $l=100$, and $\vartheta=10$.

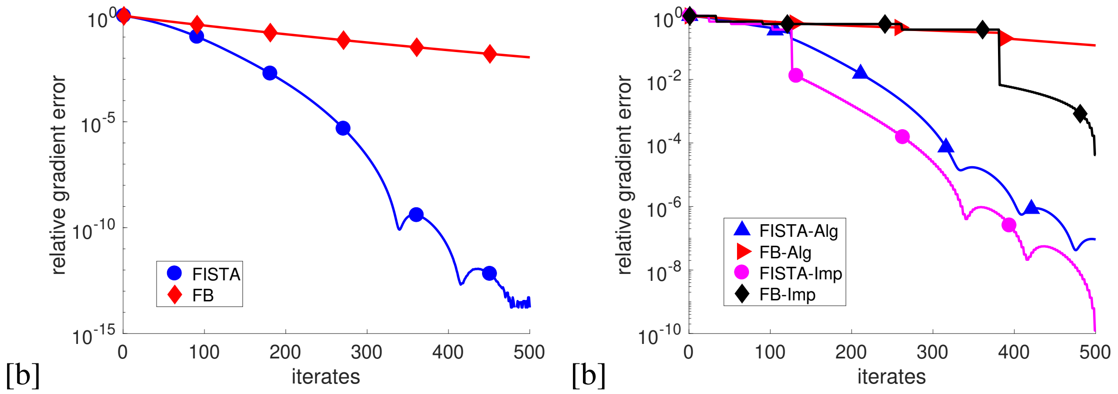


Fig. 7.3 Left: The relative objective of (7.36) with respect to the iterates, for FISTA and FB. Right: The relative error $\|Dx^k(\vartheta) - Dx(\vartheta)\| / \|Dx^1(\vartheta) - Dx(\vartheta)\|$ with respect to the iterate, for FISTA and FB, and implicit (Imp) and algorithmic (Alg) differentiation.

See Figure 7.3 for the results. We observe that the implicit derivative from the FISTA method converges to the true derivative the fastest, with the algorithmic derivative from FISTA slightly behind.

7.7 Conclusion and outlook

In this chapter, we have studied the differentiability of solution mappings of parametrised variational problems under assumptions of partial smoothness. We have obtained new results on the piecewise differentiability of the solution mapping, and thereby characterised the Clarke subdifferential of the solution mapping $\overline{E}(\vartheta)$. Furthermore, we have studied algorithmic differentiation, showing that the same analysis can lead to convergence guarantees to the implicit derivative.

These results open the door for future applications of bilevel optimisation of nonsmooth variational problems. Future work will be dedicated to studying bilevel optimisation for saddle-point problems and algorithmic differentiation of primal-dual methods (see Section 8.3.3), its application to dictionary learning and to learning sampling patterns for MRI, along the lines of [205]. Furthermore, we want to investigate how the combination of algorithmic differentiation and stopping rules affect the stability of schemes for bilevel optimisation (see Section 8.3.3).

Chapter 8

Summary, discussion, and outlook

8.1 Summary

In this thesis, we studied discrete gradient methods from geometric numerical integration for solving various classes of optimisation problems. These methods preserve structures of differential systems, including the dissipative structure of gradient flows and the inverse scale space flow.

In the context of optimisation of continuously differentiable functions, we prove convergence rates of $O(1/k)$ for L-smooth, convex functions, and linear rates for L-smooth functions that satisfy the Polyak–Łojasiewicz inequality. Furthermore, we prove that the discrete gradient method is well-defined for all time steps $\tau_k > 0$. Finally we propose a scheme for solving the discrete gradient equation, which we demonstrate is superior in stability and efficiency for different optimisation problems.

Furthermore, the Itoh–Abe discrete gradient is derivative-free, thereby providing a notion of gradient flow-type dissipation in a derivative-free setting. With this in mind, we have studied derivative-free discrete gradient methods applied to locally Lipschitz continuous functions, and proven that the method is well-defined and converges to a set of stationary points in the nonsmooth, nonconvex Clarke subdifferential framework.

A central motivation for studying derivative-free, nonsmooth optimisation is bilevel optimisation problems, in which the parameters of a nonsmooth variational optimisation problem are optimised with respect to some higher-level cost function. The reason for this is that these problems are nonconvex, nonsmooth, and their differential structure is far from trivial, and can therefore naturally be treated in a black-box optimisation setting.

Black-box problems, including bilevel problems, often include parameter constraints, ranging from a nonnegativity criterion to complicated, implicitly defined constraints. We have therefore extended the Itoh–Abe discrete gradient method to a nonsmooth, nonconvex, constrained optimisation setting, and proven that if the constraint is epi-Lipschitzian (i.e. the level set of a locally Lipschitz continuous function), then the method converges to a set of constrained Clarke stationary points.

Furthermore, we have extended the derivative-free optimisation framework to solving the inverse scale space flow, which allows us to incorporate additional structure into the iterative procedure, such as promoting sparsity in the reconstruction. This enabled us to modify established schemes such as Gauss-Seidel and SOR and achieve significantly faster convergence.

In the final part of the thesis, we returned to bilevel optimisation problems, and investigated how one can differentiate the solution map corresponding to nonsmooth variational problems, provided that there is a partly smooth structure to exploit. Building on this, we show that the solution map is piecewise differentiable, and furthermore that we can differentiate the iterates of optimisation algorithms with provable convergence guarantees to the true gradient. This opens the door for future applications for example in dictionary learning.

8.2 Discussion

Numerical integration has been central to many important innovations in mathematical optimisation in the last decade, and it will continue to play a central role in the years to come. A recurring advantage we observed through this thesis is that by using numerical methods that preserve some structure, one can relax other assumptions of the optimisation problem while retaining this structure. For example, the dissipative structure of a gradient flow is preserved for discrete gradient methods, even when the differential structure of the objective function is lost, yielding a method with linear convergence rates for strongly convex, smooth problems, while also having convergence guarantees in the Clarke subdifferential framework for locally Lipschitz continuous functions.

Furthermore, as the work on discrete gradient methods for the inverse scale space flow demonstrates, discrete gradient methods can preserve dissipative structures beyond those based on the Euclidean metric. This suggests that we can apply such methods to more complicated, dissipative flows.

Finally, we note that for bilevel optimisation problems, one can either treat it as a black-box problem and apply Itoh–Abe type methods, or one can seek to exploit its partly smooth structure to compute gradients, implicitly or algorithmically. These two differing approaches to an optimisation problem demonstrate somehing more general about this research area, namely that to tackle a challenging problem, one can either seek to develop new methods that are implementable based on the current knowledge of the problem, or one can seek to analyse the problem further, to understand its structure better.

8.3 Outlook

For the rest of this chapter, we discuss future directions of research, building on the work presented in this thesis.

8.3.1 Discrete gradient methods for solving Wasserstein gradient flow

Building on the idea in Chapter 6 that discrete gradient methods can preserve dissipative structures beyond the gradient flow in the Euclidean setting, we consider the use of discrete gradient methods for solving *Wasserstein gradient flow*.

The Wasserstein distance is a distance on the space of probability measures. It is colloquially referred to as the *earth mover's distance*, as if one considers two piles of dirt as the probability measures, the Wasserstein distance represents the minimal amount of work required to turn one pile into the other. This notion of distance is inherently different from the Euclidean distance, and is of great importance to many scientific and mathematical fields, including economics concerning optimal resource allocation, probability theory and the geometric structures of measures, and computer vision.

A motivation for considering discrete gradient methods for the Wasserstein gradient flow is that solving this flow tends to be significantly computationally intensive, and one needs to ensure the preservation of properties such as nonnegativity and preservation of mass, so that each iterate remains a probability distribution.

We now provide a brief overview of the Wasserstein distance and its corresponding gradient flow, and consider the application of discrete gradient methods. However, we leave out several details on measure theory, and simply point out that there is a rich mathematical theory and historical background to Wasserstein distances and optimal transport—see e.g. [5, 176, 199, 216] for further information.

For a separable metric space $(X, d : X \times X \to [0, \infty))$, denote by $\mathscr{B}(X)$ the family of Borel subsets of X and by $\mathscr{P}(X)$ the sets of probability measures. Furthermore, for $p > 0$, denote by $\mathscr{P}_p(X)$ the set of probability measures with finite p-th moment, i.e. such that

$$\int_X d(x,y) d\mu(x) < \infty \quad \text{for some } y \in X.$$

Given two separable metric spaces X and Y, a measure $\mu \in \mathscr{P}(X)$, and a μ-measurable function $r : X \to Y$, the *push-forward of μ through r* is the measure $r_{\#}\mu \in \mathscr{P}(Y)$ defined as

$$r_{\#}\mu(B) := \mu(r^{-1}(B)), \quad \text{for all } B \in \mathscr{B}(Y).$$

Consider the product space $X \times Y$, and denote by π^1 and π^2 the projection operators on $X \times Y$ onto X and Y respectively. For $\mu^1 \in \mathscr{P}(X)$ and $\mu^2 \in \mathscr{P}(Y)$, the class of *multiple plans* with marginals μ^i, $i \in \{1,2\}$ is given by

$$\Gamma(\mu^1,\mu^2) := \left\{ \mu \in \mathscr{P}(X \times Y) \ : \ \pi^i_{\#}\mu = \mu^i, \ i = 1,2 \right\}.$$

We can now define Wasserstein distances. For our purposes, we assume that (X,d) is a separable Hilbert space so that $d(x,y) := \|x-y\|^2$, and we consider a compact, convex, nonempty subset $\Omega \subset X$. Then, given probability measures μ^1, $\mu^2 \in \mathscr{P}_2(\Omega)$, the *2nd Wasserstein distance* between μ^1 and μ^2 is defined as

$$W_2^2(\mu^1,\mu^2) := \min \left\{ \int_{\Omega\times\Omega} \|x-y\|^2 d\mu(x,y) \ : \ \mu \in \Gamma(\mu^1,\mu^2) \right\}. \tag{8.1}$$

See [5, Chapter 7] for properties of Wasserstein distances.

We briefly discuss the Wasserstein gradient flow in $\mathscr{P}_2(\Omega)$, and refer for the details to [199, Chapter 8]. Recall the formulation of the implicit gradient descent step in Chapter 1,

$$x^{k+1} = \underset{y\in\mathbb{R}^n}{\arg\min} F(y) + \frac{1}{2\tau_k}\|y - x^k\|^2.$$

For the metric space $(\mathscr{P}_2(\Omega), W_2)$ and $\rho^k \in \mathscr{P}_2(\Omega)$, the analogous scheme becomes

$$\rho^{k+1} = \underset{\rho\in\mathscr{P}_2(\Omega)}{\arg\min} \ F(y) + \frac{1}{2\tau_k} W_2^2(\rho,\rho^k), \tag{8.2}$$

for a functional $F : \mathscr{P}_2(\Omega) \to \overline{\mathbb{R}}$. This scheme is called the *minimising movement scheme*, and is a natural way of defining gradient flow-type schemes with respect to metrics that do not admit an inner product structure[1].

In the limit τ_k, this yields the differential equation

$$\dot{p}(t) - \nabla \cdot \left(\rho \nabla (\frac{\delta F}{\delta \rho}(\rho)) \right) = 0, \tag{8.3}$$

[1] Note the similar connection for Bregman iterative methods and inverse scale space flow in Chapter 6.

with homogeneous Neumann boundary conditions on $\operatorname{bd}\Omega$, i.e. $\dot{\rho}\nabla(\frac{\delta F}{\delta\rho}(\rho))\cdot\widehat{n}=0$, where $\widehat{n}$ is the outward normal vector to $\operatorname{bd}\Omega$. Here ∇ refers to the spatial gradient, and $\frac{\delta F}{\delta\rho}$ is the *first variation* of F [199, Definition 7.12], i.e. the Gateaux derivative of F. We thus propose to apply discrete gradient methods to discretise this equation.

For a functional $F:\mathscr{P}(\Omega)\to\overline{\mathbb{R}}$, one would want a discrete gradient $\overline{\nabla}F$ to be consistent with the first variation $\frac{\delta F}{\delta\rho}$ in the limit, and to satisfy a mean value property, say, of the form

$$\int_\Omega \overline{\nabla}F(\rho^1,\rho^2)d(\rho^1-\rho^2)(x)=F(\rho^1)-F(\rho^2). \tag{8.4}$$

Assuming such a discrete gradient exists, we propose the following scheme. Given a starting point $\rho^0\in\mathscr{P}_2(\Omega)$ and time steps $(\tau_k)_{k\in\mathbb{R}}$, we want to solve

$$\begin{aligned}
&\rho^{k+1}=\rho^k+\tau_k\nabla\cdot\left(\rho^k\nabla(\overline{\nabla}F(\rho^k,\rho^{k+1}))\right),\\
&\rho^{k+1}\in\mathscr{P}_2(\Omega),\quad \rho^k\nabla(\overline{\nabla}F(\rho^k,\rho^{k+1})\cdot\widehat{n}=0.
\end{aligned}\tag{8.5}$$

The second requirement above is equivalent to $\nabla\cdot\left(\rho^k\nabla(\overline{\nabla}F(\rho^k,\rho^{k+1}))\right)\in T_{\rho^k}\mathscr{P}(\Omega)$, where $T_\rho\mathscr{P}(\Omega)$ denotes the tangent space of $\mathscr{P}(\Omega)$ at ρ^k and is given by [90, 171]

$$T_\rho\mathscr{P}(\Omega)=\left\{\eta\in L^2(\Omega),:\ \eta=\nabla\cdot\rho\nabla\Phi\ \text{ with }\Phi\text{ s.t. }\rho\nabla\Phi\cdot\widehat{n}=0\text{ on }\operatorname{bd}\Omega\right\}.$$

We include this requirement to allow integration by parts with respect to the divergence term $\nabla\cdot$.

We will show that this scheme conserves both mass and the dissipative structure of the original minimising movement scheme (8.2). We first show mass conservation. For this, we simply compute

$$\begin{aligned}
\frac{\rho^{k+1}(X)-\rho^k(X)}{\tau_k}&=\frac{\int_X\mathbb{1}(x)d(\rho^{k+1}-\rho^k)(x)}{\tau_k}\\
&=\int_X\mathbb{1}(x)d\left(\nabla\cdot\left(\rho^{k+1}\nabla(\overline{\nabla}F(\rho^k,\rho^{k+1}))\right)\right)(x)\\
&=-\int_X\nabla\mathbb{1}(x)d\left(\left(\rho^k\nabla(\overline{\nabla}F(\rho^k,\rho^{k+1}))\right)\right)(x)=0,
\end{aligned}$$

where we have applied (8.5) and integration by parts, and used the fact that $\mathbb{1}$ is a constant function, so $\nabla\mathbb{1}=0$.

Next we characterise the dissipative structure of the discrete gradient scheme. Assuming that ρ^k is absolutely continuous with respect to the Lebesgue measure, the dissipative structure for the minimising movement scheme (8.2) can be expressed as

$$F(\rho^k) - F(\rho^{k+1}) \geq \frac{1}{2\tau_k} W_2^2(\rho^{k+1}, \rho^k) = \frac{\tau_k}{2} \int_X \left\| \nabla \left(\frac{\delta F}{\delta \rho}(\rho^{k+1}) \right) \right\|^2 d\rho^k(x) \tag{8.6}$$

due to [199, Theorem 1.17] and [199, Equation (8.4)].

Applying the measure space analogue of (2.12) for the discrete gradient method, we derive

$$\begin{aligned} F(\rho^k) - F(\rho^{k+1}) &= \int_X \overline{\nabla} F(\rho^k, \rho^{k+1}) d(\rho^k - \rho^{k+1})(x) \\ &= -\tau_k \int_X \overline{\nabla} F(\rho^k, \rho^{k+1}) d\left(\nabla \cdot \left(\rho^k \nabla (\overline{\nabla} F(\rho^k, \rho^{k+1})) \right) \right)(x) \\ &= \tau_k \int_X \|\nabla (\overline{\nabla} F(\rho^k, \rho^{k+1}))\|^2 d\rho^k(x). \end{aligned} \tag{8.7}$$

Here we have again applied (8.5) and used integration by parts. From this, we observe that (8.7) is a discrete gradient counterpart to (8.6). Thus the proposed discrete gradient method preserves mass and the dissipative structure.

One would also want to ensure preservation of nonnegativity of ρ^{k+1}. For the time being, it is unclear how to achieve this. Another challenge is to allow for a change in support of the measure, i.e. so that $\mathrm{supp}(\rho^{k+1}) \neq \mathrm{supp}(\rho^k)$, which in the formulation (8.5) is not possible due to the presence of ρ^k as a factor in the second term on the right-hand side.

Furthermore, the requirement that $\nabla \cdot \left(\rho^k \nabla (\overline{\nabla} F(\rho^k, \rho^{k+1}) \right) \in T_{\rho^k} \mathscr{P}(\Omega)$ might be prohibitive, and one could instead seek to employ a discrete gradient formulation for the divergence with a corresponding notion of integrating by parts. We leave these issues for future work.

8.3.2 The mean value discrete gradient for nonsmooth optimisation

A central part of this thesis has been the application of the Itoh–Abe discrete gradient method for nondifferentiable functions. In what follows, we consider the possibility of using the mean value discrete gradient

$$\overline{\nabla} F(x, y) = \int_0^1 \nabla F(sx + (1-s)y) \, \mathrm{d}s$$

when F is nondifferentiable.

There are reasons why this is of interest, beyond merely extending geometric numerical integration concepts to the nonsmooth setting. Unlike the Itoh–Abe discrete gradient, the mean value discrete gradient preserves additional first-order characteristics of the function, such as Lipschitz continuity and strong monotonicity cf. Proposition 3.7. This in turn can allow us to prove further properties for discrete gradient methods applied to gradient flows and inverse scale space flow. Consider for example the regularisation properties that can be proven for Bregman and linearised Bregman iterations, e.g. Fejér monotonicity [18, Lemmas 6.4, 6.11] and convergence to J-minimising solutions [18, Lemmas 6.7, 6.13]. Such results seem to not hold for the Itoh–Abe discrete gradient method because in this case $\overline{\nabla}F$ does not inherit crucial properties of ∇F.

We propose to define a mean value discrete gradient for a nonsmooth function $F : \mathbb{R}^n \to \mathbb{R}$ accordingly. For $x, y \in \mathbb{R}^n$, we choose

$$\overline{\nabla}F(x,y) = \int_0^1 p(s)\,\mathrm{d}s, \quad p(s) \in \partial F(sx + (1-s)y),$$

provided such a subgradient selection $p(s)$ is integrable. Of course the discrete gradient is no longer unique, as subgradients are not unique. However, one could still seek to prove that the mean value and consistency properties of discrete gradients (2.12)-(2.6) hold for any such discrete gradient representation.

To show the mean value property, we could proceed as follows. If F satisfies the assumptions of partial smoothness in Chapter 7, namely Assumption 7.23, then we know that $\mathrm{par}(\partial F(x)) = N_x\mathcal{M}$ whenever F is partly smooth at x relative to $\mathcal{M}$. Therefore, if any segment $\{sx + (1-s)y \ : \ s \in [a,b]\}$ is contained in $\mathcal{M}$, then $x - y \in T_{sx+(1-s)y}\mathcal{M}$ for $s \in [a,b]$, and therefore, for any selection $p(s) \in \partial F(xs + (1-s)y)$, $s \in [a,b]$, we have

$$\int_a^b \langle p(s), xs + (1-s)y\rangle = \int_a^b \langle \nabla F_{\mathcal{M}}(sx + (1-s)y), sx + (1-s)y\rangle,$$

by Proposition 7.19. Therefore, for any subsegment of $[x,y]$ belonging to a manifold $\mathcal{M}$ relative to which F is partly smooth, the mean value property holds. Since we assume that we can partition a bounded set into a finite number of smooth manifolds, relative to each of which F is partly smooth, one could expect that the entire segment $[x,y]$ can be partitioned into such subsegments on which the mean value property holds.

To show the consistency property, i.e. that if $x^k, y^k \to x \in \mathbb{R}^n$ and $\overline{\nabla}F(x^k, y^k) \to d$, then $d = \partial F(x)$, we can use the facts that $x \mapsto \partial F(x)$ is outer semicontinuous, and that $\partial F(x)$ is convex.

If we then wanted to show that the discrete gradient equation is well-defined

$$y = x - \tau\overline{\nabla}F(x,y),$$

we would want to generalise the existence proof Theorem 3.4 which is based on Brouwer's fixed point theorem Proposition 3.3. For this we could perhaps invoke Himmelberg's fixed point theorem which generalises that of Brouwer's to set-valued maps..

Proposition 8.1 (Himmelberg's fixed point theorem)**.** *Let C be a nonempty convex subset of a separated locally convex space $\mathcal{X}$. Let $F : \mathcal{X} \to \mathcal{X}$ be an outer semicontinuous set-valued map such that $F(x)$ is closed and convex for all $x \in C$ and such that $F(C) \subseteq C$. Then F has a fixed point.*

Provided we can show that the set of mean value discrete gradients is closed, convex and bounded, which could follow from convexity and outer semicontinuity of $\partial F(x)$, then existence would follow.

Remark 8.2. *While the Gonzalez discrete gradient would satisfy the mean value property for nonsmooth functions, it does not seem to guarantee the consistency property, i.e. outer semicontinuity. This would pose problems in terms of ensuring that $x^k \to x^*$ and $\overline{\nabla} V(x^k, x^{k+1}) \to 0$ implies $0 \in \partial V(x^*)$.*

From here, we can prove under mild conditions on F that the limit set of iterates converge to stationary points, i.e. points x^* such that $0 \in \partial F(x^*)$.

Example

We consider the basis pursuit denoising problem

$$F(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1. \tag{8.8}$$

The set of mean value discrete gradients of $\|x\|_1$ is given by

$$\left[\overline{\nabla}\|\cdot\|_1(x,y)\right]_i = \begin{cases} \mathrm{SGN}\left(\frac{x+y}{2}\right), & \text{if } \mathrm{SGN}(x_i) \cap \mathrm{SGN}(y_i) \neq \emptyset, \\ \{\frac{x_i+y_i}{|x_i-y_i|}\}, & \text{else.} \end{cases}$$

Here $\mathrm{SGN}(x)$ denotes the set-valued sign function

$$\mathrm{SGN}(x) := \begin{cases} \{1\}, & \text{if } x > 0, \\ \{-1\}, & \text{if } x < 0, \\ [0,1], & \text{if } x = 0, \end{cases}$$

or equivalently the subdifferential of the scalar absolute value function $|x|$.

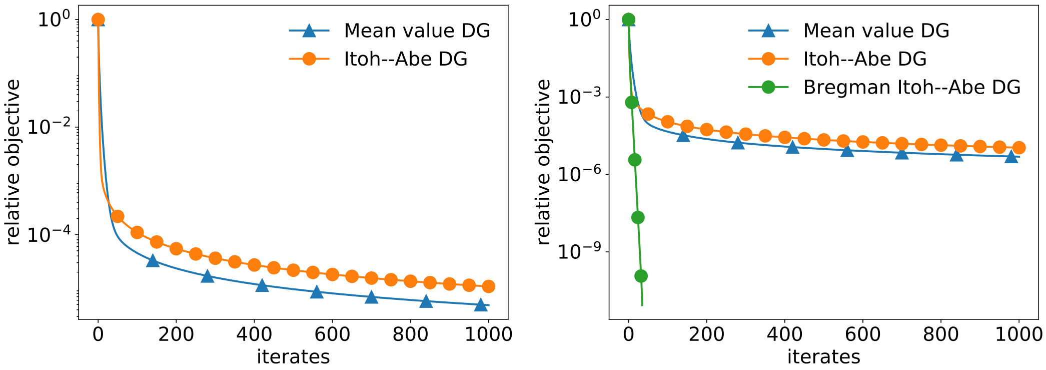


Fig. 8.1 Discrete gradient methods applied to the optimisation problem (8.10). **Left**: A comparison of the mean value discrete gradient method and the Itoh–Abe method. **Right**: A comparison of the two previous methods with the Bregman Itoh–Abe method.

In this case, the discrete gradient method becomes

$$y \in x - \tau \left(A^* \left(A \frac{x+y}{2} - b \right) + \lambda \overline{\nabla} \| \cdot \|_1 (x,y). \right) \tag{8.9}$$

By the previous subsection, we know that there exists a unique update y. It can be rewritten as

$$\frac{x+y}{2} = \left(A^*A + \frac{1}{\tau} I \right)^{-1} \left(\frac{2}{\tau} x + A^* b - \lambda \overline{\nabla} \| \cdot \|_1 (x,y) \right)$$

or

$$\left(A^*A + \frac{1}{\tau} I \right) \frac{x+y}{2} = \frac{2}{\tau} x + A^* b - \lambda \overline{\nabla} \| \cdot \|_1 (x,y).$$

We came up with an ad-hoc fixed point method for solving (8.9), which converged (albeit quite slowly) most of the time. We then implemented it on the optimisation problem

$$\min_x \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1, \tag{8.10}$$

where $\lambda = 1$, the dimension $n = 50$, and A is a random Gaussian matrix. We compare the nonsmooth mean value discrete gradient method with the (nonsmooth) Itoh–Abe discrete gradient method. Then, we include a comparison with the Bregman Itoh–Abe method to demonstrate the superior efficiency of these methods for nonsmooth objective functions. See Figure 8.1 for the results.

Going forward

The next thing one would need, in order to make this method interesting / feasible, is an efficient and stable method for solving the discrete gradient equation (2.8). This amounts to solving a nonsmooth fixed point problem. If the objective function is nonconvex, coming up with a general and efficient method seems tricky / unreasonable. However, it could be interesting to see whether a method can be formulated for convex objective functions. However, investigating this could involve some time and effort, so we should first determine whether this is a problem worth pursuing. One might imagine that nonsmooth fixed point problems involving monotone operators has other applications too.

8.3.3 Differentiation for nonsmooth bilevel optimisation

Convergence of algorithmic derivatives for primal-dual methods

Future work will be dedicated to studying algorithmic differentiation of other iterative methods, including *primal-dual* methods. These methods solve, for a given parameter choice ϑ,

$$\min_{x\in\mathbb{R}^n} R(x,\vartheta)+V(x,\vartheta)+(J \overset{+}{\triangledown} G)(K(\vartheta)x,\vartheta), \tag{8.11}$$

where $R(\cdot,\vartheta), V(\cdot,\vartheta)\in\Gamma_0(\mathbb{R}^n)$, $J(\cdot,\vartheta)$, $G(\cdot,\vartheta)\in\Gamma_0(\mathbb{R}^l)$, $V(\cdot,\vartheta)$ is L_V-smooth, $G(\cdot,\vartheta)$ is $1/L_G$-convex for some $L_V, L_G > 0$, and $K:\mathbb{R}^m\to\mathbb{R}^{l\times n}$ is a matrix mapping. Here $J \overset{+}{\triangledown} G$ is the parameter-dependent *infimal convolution* of J and G,

$$(J \overset{+}{\triangledown} G)(x,\vartheta) := \inf_{y\in\mathbb{R}^l} J(x-y,\vartheta)+G(y,\vartheta).$$

As is shown in [136, Section 3.1], the primal-dual method can be expressed as a forward-backward method for solving a monotone inclusion problem. Specifically,

$$z^{k+1} = (\mathcal{V}+A)^{-1}(\mathcal{V}-B)z^k, \tag{8.12}$$

where $z^k = [x^k, y^k]^T$ and

$$\mathcal{V} = \begin{bmatrix} \frac{I_n}{\gamma_R} & -K^*(\vartheta) \\ -K(\vartheta) & \frac{I_m}{\gamma_J} \end{bmatrix}, A = \begin{bmatrix} \partial_x R(\cdot,\vartheta) & K^*(\vartheta) \\ -K(\vartheta) & \partial_x J^*(\cdot,\vartheta) \end{bmatrix}, B = \begin{bmatrix} \nabla_x V(\cdot,\vartheta) & 0 \\ 0 & \nabla_x G^*(\cdot,\vartheta) \end{bmatrix},$$

and where $\mathcal{V}$ is ν-positive definite, for $\nu = (1-\sqrt{\gamma_J\gamma_R\|K\|^2})\min\{1/\gamma_J, 1/\gamma_R\}$. Here $\gamma_R > 0$ and $\gamma_J > 0$ are the primal and dual time steps respectively.

We require the following conditions. Both $V+R$ and G^*+J^* satisfy Assumption 7.23, $\vartheta \mapsto K(\vartheta)$ is C^1-smooth, and we have the nondegeneracy assumptions

$$\begin{aligned} -K^*(\vartheta)y^* - \nabla_x V(x^*,\vartheta) &\in \operatorname{ri}\left(\partial_x R(x^*,\vartheta)\right), \\ K(\vartheta)x^* - \nabla_x G^*(y^*,\vartheta) &\in \operatorname{ri}\left(\partial_x J^*(y^*,\vartheta)\right). \end{aligned}$$

Furthermore, the time steps γ_R and γ_J satisfy

$$2\min\{\tfrac{1}{L_V},\tfrac{1}{L_G}\}\min\{\tfrac{1}{\gamma_J},\tfrac{1}{\gamma_R}\}(1-\sqrt{\gamma_J\gamma_R\|K(\vartheta)\|^2}) > 1. \tag{8.13}$$

In [136, Theorem 3.2], finite activity identification is proven under these conditions. We would like to establish that algorithmic differentiation of the primal-dual methods would ensure convergence to the correct limit.

By arguing as we did for the forward-backward methods, for sufficiently large k, the update map (8.12) is locally continuously differentiable with respect to z^k and ϑ, and we have

$$Dz^{k+1}(\vartheta) = M_k Dz^k(\vartheta) + b^k, \quad M_k \to M, \quad b^k \to b,$$

where

$$\begin{aligned} M &= (\mathcal{V}+M_A)^\dagger(\mathcal{V}-M_B), & b &= -(\mathcal{V}+M_A)^\dagger(D_A+D_B), \\ M_A &= \begin{bmatrix} \nabla^2_{\mathcal{M}_{x^*}} R(x^*,\vartheta) & K^*(\vartheta) \\ -K(\vartheta) & \nabla^2_{\mathcal{M}_{y^*}} J^*(y^*,\vartheta) \end{bmatrix}, & M_B &= \begin{bmatrix} \nabla^2_{\mathcal{M}_{x^*}} V(x^*,\vartheta) & \mathbf{0} \\ 0 & \nabla^2_{\mathcal{M}_{y^*}} G^*(y^*,\vartheta) \end{bmatrix}, \\ D_A &= \begin{bmatrix} D_\vartheta \nabla_{\mathcal{M}_{x^*}} R(x^*,\vartheta) \\ D_\vartheta \nabla_{\mathcal{M}_{y^*}} J^*(y^*,\vartheta) \end{bmatrix}, & D_B &= \begin{bmatrix} D_\vartheta \nabla_{\mathcal{M}_{x^*}} V(x^*,\vartheta) - DK^*(\vartheta)y^* \\ D_\vartheta \nabla_{\mathcal{M}_{y^*}} G^*(y^*,\vartheta) - DK(\vartheta)x^* \end{bmatrix}. \end{aligned}$$

We want to show that $\rho(M) < 1$. Suppose $z = [x,y]^T \in \mathbb{C}^{n+l}$ and $\lambda \in \mathbb{C}$ satisfy $Mz = \lambda z$. As usual, we can apply Proposition 2.8 to conclude that $\lambda \in \mathbb{R}$. It follows that $z \in T_{x^*}\mathcal{M}_{x^*} \times T_{y^*}\mathcal{M}_{y^*}$, in which case we rearrange to get

$$(1-\lambda)\mathcal{V}z = \lambda M_A z + M_B z.$$

Taking the inner product with respect to z on each side gives us

$$(1-\lambda)\langle z,\mathcal{V}z\rangle = \lambda\langle x,M_R x\rangle + \lambda\langle y,M_{J^*}y\rangle + \langle x,M_V x\rangle + \langle y,M_{G^*}y\rangle,$$

where

$$M_V = \nabla^2_{\mathcal{M}_{x^*}} V(x^*,\lambda),\ M_R = \nabla^2_{\mathcal{M}_{x^*}} R(x^*,\lambda),\ M_{J^*} = \nabla^2_{\mathcal{M}_{y^*}} J^*(y^*,\lambda),\ M_{G^*} = \nabla^2_{\mathcal{M}_{y^*}} G^*(y^*,\lambda).$$

We observe that $\lambda < 1$, since otherwise the left-hand side is nonpositive, while the right-hand side is strictly positive by the strong convexity assumptions for $V+R$ and $J^* + G^*$. Suppose $\lambda \leq -1$. Then since $\mathcal{V}$ is ν-positive-definite,

$$(1-\lambda)\langle z, \mathcal{V}z\rangle \geq 2\nu\|z\|^2.$$

On the other hand, M_R and M_{J^*} are positive-definite, and M_V and M_{G^*} are L_V- and L_G-smooth respectively, we have

$$\lambda\langle x, M_R x\rangle + \lambda\langle y, M_{J^*}y\rangle + \langle x, M_V x\rangle + \langle y, M_{G^*}y\rangle \leq L_V\|x\|^2 + L_G\|y\|^2 \leq \frac{1}{\min\{\frac{1}{L_V},\frac{1}{L_G}\}}\|z\|^2.$$

By (8.13),

$$\frac{1}{\min\{\frac{1}{L_V},\frac{1}{L_G}\}}\|z\|^2 < 2\nu\|z\|^2 \leq (1-\lambda)\langle z, \mathcal{V}z\rangle.$$

Therefore the equality cannot hold when $\lambda \leq -1$. Therefore $|\lambda| < 1$ which implies that $\rho(M) < 1$.

This suggests that one can also ensure convergence of the algorithmic derivatives for primal-dual methods under the partial smoothness framework. One drawback for this is that one requires strong convexity both in the primal and the dual component, while implicit differentiation of the solution map $x(\vartheta)$ only requires strong convexity of the primal problem formulation. A possible work-around for this is to replace the strong convexity requirement with something weaker, such as restricted injectivity [135].

One would need to show that the results of [130, Theorem 5.7] generalise to implicit differentiation with respect to optimality conditions of saddle-point problems. For this, one could potentially build on the framework in [131].

Stability of algorithmic differentiation under stopping rule

One motivation for studying algorithmic and implicit differentiation of the nonsmooth solution maps is to assess the advantages and disadvantages of choosing either approach for derivative estimation. While implicit differentiation yields the 'true' derivative of the solution map, this is only the case if the minimiser $x(\vartheta)$ has been located up to error tolerance. In many cases, the lower-level optimisation problems are ill-conditioned and high-dimensional, and it might not be feasible to compute enough iterates x^k to estimate $x(\vartheta)$ sufficiently for implicit differentiation to be reliable. Indeed, the iterates might not have located the smooth manifold containing the minimiser yet.

In such cases, algorithmic differentiation has the advantage that even if the final iterate $x^K(\vartheta)$ is a poor approximation of $x(\vartheta)$, the algorithmic derivative $Dx^K(\vartheta)$ corresponds to the derivative of the function that is being computed, namely $x^K(\vartheta) \approx \mathcal{A}^K(x^0, \vartheta)$. That is, assuming the number of iterations K are fixed.

However, it is often preferrable to employ a stopping rule to determine the number of iterations. For example, one could stop once the measure of progress $\|x^k - x^{k+1}\|/\|x^k\|$ is less than some tolerance ε, or, for primal-dual methods once the primal-dual gap is sufficiently small [50]. From the perspective of algorithmic differentation, this introduces an element of discontinuity in the relation between the parameter choice and output $x^K(\vartheta)$.

Future research will deal with the stability or instability of using algorithmic differentiation combined with stopping rules.

References

[1] Abdulle, A. (2015). Explicit stabilized runge-kutta methods. Technical report, Springer.

[2] Absil, P.-A., Mahony, R., and Andrews, B. (2005). Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547.

[3] Absil, P.-A., Mahony, R., and Trumpf, J. (2013). An extrinsic look at the Riemannian Hessian. In *International Conference on Geometric Science of Information*, pages 361–368. Springer.

[4] Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Pure and Applied Mathematics. Academic Press, Cambridge, MA, USA, 2nd edition.

[5] Ambrosio, L., Gigli, N., and Savare, G. (2008). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2nd edition.

[6] Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992). Image coding using wavelet transform. *IEEE Transactions on image processing*, 1(2):205–220.

[7] Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łsojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457.

[8] Attouch, H., Peypouquet, J., and Redont, P. (2014). A dynamical approach to an inertial forward-backward algorithm for convex minimization. *SIAM Journal on Optimization*, 24(1):232–256.

[9] Audet, C. and Dennis Jr, J. E. (2006). Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on optimization*, 17(1):188–217.

[10] Audet, C. and Hare, W. (2017). *Derivative-Free and Blackbox Optimization*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 1st edition.

[11] Aussel, D. (1998). Subdifferential properties of quasiconvex and pseudoconvex functions: unified approach. *Journal of optimization theory and applications*, 97(1):29–45.

[12] Bauschke, H. H., Combettes, P. L., et al. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer.

[13] Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

[14] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

[15] Beck, A. and Tetruashvili, L. (2013). On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060.

[16] Becker, S., Bobin, J., and Candès, E. J. (2011). NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39.

[17] Benning, M., Betcke, M. M., Ehrhardt, M. J., and Schönlieb, C.-B. (2017a). Choose your path wisely: gradient descent in a Bregman distance framework. *ArXiv e-prints*.

[18] Benning, M. and Burger, M. (2018). Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111.

[19] Benning, M., Gilboa, G., Grah, J. S., and Schönlieb, C.-B. (2017b). Learning filter functions in regularisers by minimising quotients. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 511–523. Springer.

[20] Benning, M., Riis, E. S., and Schönlieb, C.-B. (2020). Bregman Itoh–Abe methods for sparse optimisation. *J. Math. Imaging Vision*.

[21] Bertocchi, C., Chouzenoux, E., Corbineau, M.-C., Pesquet, J.-C., and Prato, M. (2018). Deep Unfolding of a Proximal Interior Point Method for Image Restoration. *arXiv e-prints*, page arXiv:1812.04276.

[22] Bertsekas, D. P. (2003). *Nonlinear Programming*. Athena Scientific, Belmont, MA, USA, 2nd edition.

[23] Betancourt, M., Jordan, M. I., and Wilson, A. C. (2018). On symplectic optimization. *arXiv e-prints*, page arXiv:1802.03653.

[24] Bolte, J., Daniilidis, A., Lewis, A., and Shiota, M. (2007). Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572.

[25] Borwein, J. M. and Zhu, Q. J. (1999). A survey of subdifferential calculus with applications. *Nonlinear Analysis: Theory, Methods & Applications*, 38(6):687–773.

[26] Bounkhel, M. and Thibault, L. (2002). On various notions of regularity of sets in nonsmooth analysis. *Nonlinear Analysis: Theory, Methods & Applications*, 48(2):223–246.

[27] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

[28] Bredies, K. and Holler, M. (2015). A TGV-based framework for variational image decompression, zooming, and reconstruction. part I: Analytics. *SIAM Journal on Imaging Sciences*, 8(4):2814–2850.

[29] Bredies, K., Kunisch, K., and Pock, T. (2010). Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526.

[30] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

[31] Brouwer, L. E. J. (1911). Über Abbildung von Mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115.

[32] Burger, M. (2016). Bregman distances in inverse problems and partial differential equations. *Advances in Mathematical Modeling, Optimization and Optimal Control*, page 3.

[33] Burger, M., Gilboa, G., Moeller, M., Eckardt, L., and Cremers, D. (2016). Spectral decompositions using one-homogeneous functionals. *SIAM Journal on Imaging Sciences*, 9(3):1374–1408.

[34] Burger, M., Gilboa, G., Osher, S., and Xu, J. (2006). Nonlinear inverse scale space methods. *Communications in Mathematical Sciences*, 4(1):179–212.

[35] Burger, M., Möller, M., Benning, M., and Osher, S. (2013). An adaptive inverse scale space method for compressed sensing. *Mathematics of Computation*, 82(281):269–299.

[36] Burger, M. and Osher, S. (2013). A guide to the TV zoo. In *Level set and PDE based reconstruction methods in imaging*, pages 1–70. Springer, Berlin.

[37] Burger, M., Resmerita, E., and He, L. (2007). Error estimation for bregman iterations and inverse scale space methods in image restoration. *Computing*, 81(2-3):109–135.

[38] Burke, J. V., Curtis, F. E., Lewis, A. S., Overton, M. L., and Simões, L. E. (2018). Gradient sampling methods for nonsmooth optimization. *arXiv e-prints*.

[39] Cai, J.-F., Osher, S., and Shen, Z. (2009). Linearized bregman iterations for compressed sensing. *Mathematics of Computation*, 78(267):1515–1536.

[40] Calatroni, L., Cao, C., De Los Reyes, J. C., Schönlieb, C.-B., and Valkonen, T. (2017). Bilevel approaches for learning of variational imaging models. In *Variational Methods*, pages 252–290. Walter de Gruyter GmbH.

[41] Candes, E., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509.

[42] Cartis, C., Fiala, J., Marteau, B., and Roberts, L. (2018). Improving the Flexibility and Robustness of Model-Based Derivative-Free Optimization Solvers. *ArXiv e-prints*.

[43] Cartis, C., Gould, N. I., and Toint, P. L. (2018). Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. *arXiv e-prints*, page arXiv:1811.01220.

[44] Celledoni, E., Eidnes, S., Owren, B., and Ringholm, T. (2018). Dissipative numerical schemes on riemannian manifolds with applications to gradient flows. *SIAM Journal on Scientific Computing*, 40(6):A3789–A3806.

[45] Celledoni, E., Grimm, V., McLachlan, R. I., McLaren, D. I., O'Neale, D., Owren, B., and Quispel, G. R. W. (2012). Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method. *Journal of Computational Physics*, 231(20):6770–6789.

[46] Censor, Y. and Zenios, S. A. (1992). Proximal minimization algorithm with d-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464.

[47] Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1-2):89–97.

[48] Chambolle, A., Caselles, V., Cremers, D., Novaga, M., and Pock, T. (2010). *An Introduction to Total Variation in image analysis*. Radon Series on Computational and Applied Mathematics. De Gruyter.

[49] Chambolle, A. and Dossal, C. (2015). On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization theory and Applications*, 166(3):968–982.

[50] Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145.

[51] Chavel, I. (2006). *Riemannian Geometry: A Modern Introduction*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition.

[52] Chen, Y., Ranftl, R., and Pock, T. (2014). Insights into analysis operator learning: From patch-based sparse models to higher order MRFs. *IEEE Transactions on Image Processing*, 23(3):1060–1072.

[53] Clarke, F. H. (1973). *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*. PhD thesis, University of Washington.

[54] Clarke, F. H. (1990). *Optimization and Nonsmooth Analysis*. Classics in Applied Mathematics. SIAM, Philadelphia, 1st edition.

[55] Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer.

[56] Combettes, P. L. and Pesquet, J.-C. (2018). Deep neural network structures solving variational inequalities. *arXiv e-prints*, page arXiv:1808.07526.

[57] Combettes, P. L. and Wajs, V. R. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200.

[58] Csiba, D. and Richtárik, P. (2017). Global Convergence of Arbitrary-Block Gradient Methods for Generalized Polyak–Łojasiewicz Functions. *ArXiv e-prints*.

[59] Curtis, F. E. and Que, X. (2013). An adaptive gradient sampling algorithm for non-smooth optimization. *Optimization Methods and Software*, 28(6):1302–1324.

[60] Curtis, F. E. and Que, X. (2015). A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees. *Mathematical Programming Computation*, 7(4):399–428.

[61] Czarnecki, M.-O. and Gudovich, A. N. (2010). Representations of epi-Lipschitzian sets. *Nonlinear Analysis: Theory, Methods & Applications*, 73(8):2361–2367.

[62] Daniilidis, A., Drusvyatskiy, D., and Lewis, A. S. (2014). Orthogonal invariance and identifiability. *SIAM Journal on Matrix Analysis and Applications*, 35(2):580–598.

[63] Daniilidis, A., Hare, W., and Malick, J. (2006). Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. *Optimization*, 55(5-6):481–503.

[64] Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61. Siam.

[65] De los Reyes, J. C., Schönlieb, C.-B., and Valkonen, T. (2016). The structure of optimal parameters for image restoration problems. *Journal of Mathematical Analysis and Applications*, 434(1):464–500.

[66] De los Reyes, J. C., Schönlieb, C.-B., and Valkonen, T. (2017). Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, 57(1):1–25.

[67] Deledalle, C.-A., Papadakis, N., Salmon, J., and Vaiter, S. (2017). CLEAR: Covariant LEAst-Square Refitting with applications to image restoration. *SIAM Journal on Imaging Sciences*, 10(1):243–284.

[68] Deledalle, C.-A., Vaiter, S., Fadili, J., and Peyré, G. (2014). Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487.

[69] Dempe, S. and Zemkoho, A. B. (2014). Kkt reformulation and necessary conditions for optimality in nonsmooth bilevel optimization. *SIAM Journal on Optimization*, 24(4):1639–1669.

[70] Deng, X. (1998). Complexity issues in bilevel linear programming. In *Multilevel optimization: Algorithms and applications*, pages 149–164. Springer.

[71] Digabel, S. L. and Wild, S. M. (2015). A taxonomy of constraints in simulation-based optimization. *arXiv e-prints*, page arXiv:1505.07881.

[72] Dong, B., Mao, Y., Osher, S., and Yin, W. (2010). Fast linearized Bregman iteration for compressive sensing and sparse denoising. *Communications in Mathematical Sciences*, 8(1):93–111.

[73] Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

[74] Dontchev, A. and Rockafellar, R. (2009). *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer Monographs in Mathematics. Springer New York.

[75] Douglas, J. and Rachford, H. H. (1956). On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439.

[76] Drusvyatskiy, D., Ioffe, A. D., and Lewis, A. S. (2016). Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization*, 26(1):513–534.

[77] DuPont, B. and Cagan, J. (2016). A hybrid extended pattern search/genetic algorithm for multi-stage wind farm optimization. *Optimization and Engineering*, 17(1):77–103.

[78] Eckstein, J. (1993). Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226.

[79] Eckstein, J. and Bertsekas, D. P. (1992). On the Douglas?Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318.

[80] Eftekhari, A., Vandereycken, B., Vilmart, G., and Zygalakis, K. C. (2018). Explicit Stabilised Gradient Descent for Faster Strongly Convex Optimisation. *ArXiv e-prints*.

[81] Ehrhardt, M. J., Riis, E. S., Ringholm, T., and Schönlieb, C.-B. (2018). A geometric integration approach to smooth optimisation: Foundations of the discrete gradient method. *ArXiv e-prints*.

[82] Ekeland, I. and Téman, R. (1999). *Convex Analysis and Variational Problems*. SIAM, Philadelphia, PA, USA, 1st edition.

[83] Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems. Mathematics and Its Applications*, volume 375. Kluwer Academic, Dordrecht, 1st edition.

[84] Evans, L. and Society, A. M. (1998). *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society.

[85] Facchinei, F. and Pang, J. (2003). *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research and Financial Engineering. Springer New York.

[86] Fadili, J., Malick, J., and Peyré, G. (2018). Sensitivity analysis for mirror-stratifiable convex functions. *SIAM Journal on Optimization*, 28(4):2975–3000.

[87] Fasano, G., Liuzzi, G., Lucidi, S., and Rinaldi, F. (2014). A linesearch-based derivative-free approach for nonsmooth constrained optimization. *SIAM Journal on Optimization*, 24(3):959–992.

[88] Fehrenbach, J., Nikolova, M., Steidl, G., and Weiss, P. (2015). Bilevel image denoising using gaussianity tests. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 117–128. Springer.

[89] Fercoq, O. and Richtárik, P. (2015). Accelerated, Parallel and Proximal Coordinate Descent. *SIAM Journal on Optimization*, 25(4):1997–2023.

[90] Feyeux, N., Vidard, A., and Nodet, M. (2018). Optimal transport for variational data assimilation. *Nonlinear Processes in Geophysics*, 25(1):55–66.

[91] Fowler, K. R., Reese, J. P., Kees, C. E., Dennis Jr, J., Kelley, C. T., Miller, C. T., Audet, C., Booker, A. J., Couture, G., Darwin, R. W., et al. (2008). Comparison of derivative-free optimization methods for groundwater supply and hydraulic capture community problems. *Advances in Water Resources*, 31(5):743–757.

[92] França, G., Sulam, J., Robinson, D. P., and Vidal, R. (2019). Conformal symplectic and relativistic optimization. *arXiv e-prints*, page arXiv:1903.04100.

[93] Gabay, D. (1983). Chapter ix applications of the method of multipliers to variational inequalities. In *Studies in mathematics and its applications*, volume 15, pages 299–331. Elsevier.

[94] Gilboa, G., Moeller, M., and Burger, M. (2016). Nonlinear spectral analysis via one-homogeneous functionals: Overview and future prospects. *Journal of Mathematical Imaging and Vision*, 56(2):300–319.

[95] Giles, J. R. (1999). A survey of Clarke's subdifferential and the differentiability of locally Lipschitz functions. In *Progress in Optimization*, pages 3–26. Springer, Boston.

[96] Goldstein, T. and Osher, S. (2009). The split Bregman method for L1-regularized problems. *SIAM journal on imaging sciences*, 2(2):323–343.

[97] Gonzalez, O. (1996). Time integration and discrete Hamiltonian systems. *Journal of Nonlinear Science*, 6(5):449–467.

[98] Gray, G. A., Kolda, T. G., Sale, K., and Young, M. M. (2004). Optimizing an empirical scoring function for transmembrane protein structure determination. *INFORMS Journal on Computing*, 16(4):406–418.

[99] Griewank, A. and Walther, A. (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition.

[100] Grimm, V., McLachlan, R. I., McLaren, D. I., Quispel, G. R. W., and Schönlieb, C.-B. (2017). Discrete gradient methods for solving variational image regularisation models. *Journal of Physics A: Mathematical and Theoretical*, 50(29):295201.

[101] Güler, O. (1991). On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419.

[102] Gürbüzbalaban, M. and Overton, M. L. (2012). On Nesterov's nonsmooth Chebyshev–Rosenbrock functions. *Nonlinear Analysis: Theory, Methods & Applications*, 75(3):1282–1289.

[103] Gürbüzbalaban, M., Ozdaglar, A., Parrilo, P. A., and Vanli, N. (2017). When cyclic coordinate descent outperforms randomized coordinate descent. In *Advances in Neural Information Processing Systems*, pages 6999–7007.

[104] Hairer, E. and Lubich, C. (2013). Energy-diminishing integration of gradient systems. *IMA Journal of Numerical Analysis*, 34(2):452–461.

[105] Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, Berlin, 2nd edition.

[106] Harten, A., Lax, P. D., and Leer, B. v. (1983). On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM review*, 25(1):35–61.

[107] Heath, M. T. (2002). *Scientific Computing: An Introductory Survey*. McGraw-Hill, New York, 1st edition.

[108] Hernández-Solano, Y., Atencia, M., Joya, G., and Sandoval, F. (2015). A discrete gradient method to enhance the numerical behaviour of Hopfield networks. *Neurocomputing*, 164:45–55.

[109] Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*. SIAM, Philadelphia, USA, 2nd edition.

[110] Hintermüller, M. (2001). A proximal bundle method based on approximate subgradients. *Computational Optimization and Applications*, 20(3):245–266.

[111] Hintermüller, M. and Wu, T. (2015). Bilevel optimization for calibrating point spread functions in blind deconvolution. *Inverse Problems & Imaging*, 9(4):1139–1169.

[112] Hiriart-Urruty, J.-B. and Lemaréchal, C. (1993). *Convex analysis and minimization algorithms I: Fundamentals*, volume 305 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathemati- cal Sciences]*. Springer, Berlin, 2nd edition.

[113] Horesh, L. and Haber, E. (2009). Sensitivity computation of the ℓ_1 minimization problem and its application to dictionary design of ill-posed problems. *Inverse Problems*, 25(9):095009.

[114] Iserles, A., Crighton, D., Ablowitz, M., Davis, S., Hinch, E., Ockendon, J., and Olver, P. (1996). *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press.

[115] Iserles, A. and Quispel, G. (2018). Why geometric numerical integration? In *Discrete Mechanics, Geometric Integration and Lie–Butcher Series*, pages 1–28. Springer.

[116] Itoh, T. and Abe, K. (1988). Hamiltonian-conserving discrete canonical equations based on variational difference quotients. *Journal of Computational Physics*, 76(1):85–102.

[117] Jaffard, S., Meyer, Y., and Ryan, R. D. (2001). *Wavelets: tools for science and technology*, volume 69. Siam.

[118] Jahn, J. (2007). *Introduction to the Theory of Nonlinear Optimization*. Springer Publishing Company, Incorporated, Berlin, 3rd edition.

[119] Kämpf, J. H. and Robinson, D. (2010). Optimisation of building form for solar energy utilisation using constrained evolutionary algorithms. *Energy and Buildings*, 42(6):807–814.

[120] Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

[121] Kiwiel, K. C. (1985). *Methods of descent for nondifferentiable optimization*, volume 1133. Springer, Berlin, 1st edition.

[122] Kiwiel, K. C. (1997). Proximal minimization methods with generalized Bregman functions. *SIAM journal on control and optimization*, 35(4):1142–1168.

[123] Kiwiel, K. C. (2010). A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 20(4):1983–1994.

[124] Kreyszig, E. (1978). *Introductory functional analysis with applications*, volume 1. Wiley, New York.

[125] Kunisch, K. and Pock, T. (2013). A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983.

[126] Kurdyka, K. (1998). On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783.

[127] Le Digabel, S. (2011). Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. *ACM Transactions on Mathematical Software*, 37(4):1–15.

[128] Le Digabel, S., Tribes, C., and Audet, C. (2009). NOMAD user guide. technical report g-2009-37. Technical report, Les cahiers du GERAD.

[129] Lee, S.-I., Lee, H., Abbeel, P., and Ng, A. Y. (2006). Efficient l1 regularized logistic regression. In *FILL IN*, Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06).

[130] Lewis, A. S. (2002). Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725.

[131] Lewis, A. S. and Liang, J. (2018). Partial smoothness and constant rank. *arXiv e-prints*, page arXiv:1807.03134.

[132] Lewis, A. S. and Malick, J. (2008). Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234.

[133] Lewis, A. S. and Overton, M. L. (2013). Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141:135–163.

[134] Liang, J., Fadili, J., and Peyré, G. (2014). Local linear convergence of forward–backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pages 1970–1978.

[135] Liang, J., Fadili, J., and Peyré, G. (2017). Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437.

[136] Liang, J., Fadili, J., and Peyré, G. (2018). Local linear convergence analysis of primal–dual splitting methods. *Optimization*, 67(6):821–853.

[137] Liang, J., Fadili, J., Peyré, G., and Luke, R. (2015). Activity identification and local linear convergence of Douglas–Rachford/ADMM under partial smoothness. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 642–653. Springer.

[138] Lions, P.-L. and Mercier, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979.

[139] Liuzzi, G. and Truemper, K. (2018). Parallelized hybrid optimization methods for nonsmooth problems using NOMAD and linesearch. *Computational and Applied Mathematics*, 37:3172–3207.

[140] Loh, P.-R., Baym, M., and Berger, B. (2012). Compressive genomics. *Nature biotechnology*, 30(7):627.

[141] Lorenz, D. A., Schöpfer, F., and Wenger, S. (2014a). The linearized Bregman method via split feasibility problems: analysis and generalizations. *SIAM Journal on Imaging Sciences*, 7(2):1237–1262.

[142] Lorenz, D. A., Wenger, S., Schöpfer, F., and Magnor, M. (2014b). A sparse Kaczmarz solver and a linearized Bregman method for online compressed sensing. *arXiv e-prints*, page arXiv:1403.7543.

[143] Lustig, M., Donoho, D. L., Santos, J. M., and Pauly, J. M. (2008). Compressed sensing MRI. *IEEE signal processing magazine*, 25(2):72.

[144] Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. (2018). Hamiltonian descent methods. *arXiv e-prints*, page arXiv:1809.05042.

[145] Mairal, J., Bach, F., and Ponce, J. (2012). Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):791–804.

[146] Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th International Conference on Computer Vision*, volume 2, pages 416–423. IEEE.

[147] McLachlan, R. I. and Quispel, G. R. W. (2001). *Six lectures on the geometric integration of ODEs*, page 155–210. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge.

[148] McLachlan, R. I., Quispel, G. R. W., and Robidoux, N. (1999). Geometric integration using discrete gradients. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 357(1754):1021–1045.

[149] Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*, volume 71. SIAM, Philadelphia, PA, USA, 1st edition.

[150] Meyer, Y. (1992). *Wavelets and operators*, volume 1. Cambridge university press.

[151] Meyer, Y. (1993). Wavelets-algorithms and applications. *Wavelets-Algorithms and applications Society for Industrial and Applied Mathematics Translation., 142 p.*

[152] Michel, P. and Penot, J.-P. (1984). Calcul sous-différentiel pour des fonctions lipschitziennes et non lipschitziennes. *C. R. Acad. Sci. Paris*, 1:269–272.

[153] Miyatake, Y., Sogabe, T., and Zhang, S.-L. (2018). On the equivalence between SOR-type methods for linear systems and the discrete gradient methods for gradient systems. *Journal of Computational and Applied Mathematics*, 342:58–69.

[154] Necoara, I., Nesterov, Y., and Glineur, F. (2018). Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39.

[155] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

[156] Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley, New York.

[157] Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376.

[158] Nesterov, Y. (2004). *Introductory Lectures on Convex Programming: A Basic Course*. Springer, New York, 1st edition.

[159] Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362.

[160] Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566.

[161] Nestruev, J. (2003). *Smooth manifolds and observables*, volume 220. Springer Science & Business Media, Berlin, 1st edition.

[162] Nocedal, J. and Wright, S. (1999). *Numerical optimization*. Springer series in operations research. Springer-Verlag, New York, 1st edition.

[163] Noll, D. (2014). Convergence of non-smooth descent methods using the Kurdyka–Łojasiewicz inequality. *Journal of Optimization Theory and Applications*, 160(2):553–572.

[164] Norton, R. A. and Quispel, G. R. W. (2014). Discrete gradient methods for preserving a first integral of an ordinary differential equation. *Discrete & Continuous Dynamical Systems - A*, 34:1147–1170.

[165] Ochs, P., Chen, Y., Brox, T., and Pock, T. (2014). ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419.

[166] Ochs, P., Ranftl, R., Brox, T., and Pock, T. (2015). Bilevel optimization with nonsmooth lower level problems. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 654–665. Springer.

[167] Ochs, P., Ranftl, R., Brox, T., and Pock, T. (2016). Techniques for gradient-based bilevel optimization with non-smooth lower level problems. *Journal of Mathematical Imaging and Vision*, 56(2):175–194.

[168] Oeuvray, R. and Bierlaire, M. (2007). A new derivative-free algorithm for the medical image registration problem. *International Journal of Modelling and Simulation*, 27(2):115–124.

[169] Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607.

[170] Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W. (2005). An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489.

[171] Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26:101–174.

[172] Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239.

[173] Passty, G. B. (1979). Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390.

[174] Pence, W., Seaman, R., and White, R. (2009). Lossless astronomical image compression and the effects of noise. *Publications of the Astronomical Society of the Pacific*, 121(878):414–427.

[175] Penot, J.-P. and Quang, P. H. (1997). Generalized convexity of functions and generalized monotonicity of set-valued maps. *Journal of Optimization Theory and Applications*, 92(2):343–356.

[176] Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

[177] Peyré, G. and Fadili, J. M. (2011). Learning analysis sparsity priors. In *Sampta'11*, pages 4–pp.

[178] Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653.

[179] Polyak, B. T. (1987). *Introduction to Optimization*. Optimization Software, Inc., New York, 1st edition.

[180] Powell, M. J. D. (2006). The NEWUOA software for unconstrained optimization without derivatives. In *Large-scale nonlinear optimization*, pages 255–297. Springer, Boston, 1st edition.

[181] Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report, University of Cambridge.

[182] Qu, Z. and Richtárik, P. (2015). Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling. In *Advances in Neural Information Processing Systems*, volume 28, pages 865—-873.

[183] Quispel, G. R. W. and Turner, G. S. (1996). Discrete gradient methods for solving ODEs numerically while preserving a first integral. *Journal of Physics A: Mathematical and General*, 29(13):341–349.

[184] Riis, E. S., Ehrhardt, M. J., Quispel, G. R. W., and Schönlieb, C.-B. (2018). A geometric integration approach to nonsmooth, nonconvex optimisation. *ArXiv e-prints*.

[185] Ringholm, T., Lazic, J., and Schonlieb, C.-B. (2018). Variational image regularization with Euler's elastica using a discrete gradient scheme. *SIAM Journal on Imaging Sciences*, 11(4):2665–2691.

[186] Robinson, S. M. (1980). Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62.

[187] Robinson, S. M. (1991). An implicit-function theorem for a class of nonsmooth functions. *Mathematics of operations research*, 16(2):292–309.

[188] Rockafellar, R. (1979a). Clarke's tangent cones and the boundaries of closed sets in $\mathbb{R}^n$. *Nonlinear analysis. Theory, methods & Applications*, 3:145–154.

[189] Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898.

[190] Rockafellar, R. T. (1979b). Directionally Lipschitzian functions and subdifferential calculus. *Proceedings of the London Mathematical Society*, 3(2):331–355.

[191] Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

[192] Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

[193] Rohde, C. A. (1965). Generalized inverses of partitioned matrices. *Journal of the Society for Industrial and Applied Mathematics*, 13(4):1033–1035.

[194] Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184.

[195] Roth, S. and Black, M. J. (2009). Fields of experts. *International Journal of Computer Vision*, 82(2):205.

[196] Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268.

[197] Rudin, W. (1976). *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, New York, 3rd edition.

[198] Rudin, W. (1987). *Real and Complex Analysis*. Higher Mathematics Series. McGraw-Hill Book Company, Singapore, 3 edition.

[199] Santambrogio, F. (2015). Optimal transport for applied mathematicians: Calculus of variations, pdes, and modeling.

[200] Santambrogio, F. (2017). $\{$Euclidean, metric, and Wasserstein$\}$ gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154.

[201] Scherzer, O. and Groetsch, C. (2001). Inverse scale space theory for inverse problems. In *International Conference on Scale-Space Theories in Computer Vision*, pages 317–325. Springer.

[202] Schmidt, M. F., Benning, M., and Schönlieb, C.-B. (2018). Inverse scale space decomposition. *Inverse Problems*, 34(4):179–212.

[203] Schöpfer, F. and Lorenz, D. A. (2018). Linear convergence of the randomized sparse Kaczmarz method. *Mathematical Programming*, pages 1–28.

[204] Shapiro, A. (1990). On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487.

[205] Sherry, F., Benning, M., Reyes, J. C. D. l., Graves, M. J., Maierhofer, G., Williams, G., Schönlieb, C.-B., and Ehrhardt, M. J. (2019). Learning the sampling pattern for MRI. *arXiv e-prints*, page arXiv:1906.08754.

[206] Stich, S. U. (2014). *Convex optimization with random pursuit*. PhD thesis, ETH Zurich.

[207] Stuart, A. and Humphries, A. R. (1996). *Dynamical systems and numerical analysis*. Cambridge University Press, Cambridge, 1st edition.

[208] Su, W., Boyd, S., and Candes, E. J. (2016). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43.

[209] Sun, R. and Ye, Y. (2016). Worst-case complexity of cyclic coordinate descent: $o(n^2)$ gap with randomized version. *ArXiv e-prints*.

[210] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147.

[211] Teboulle, M. (1992). Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690.

[212] Teboulle, M. (2018). A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96.

[213] Tseng, P. (2010). Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295.

[214] Vaiter, S., Deledalle, C., Fadili, J., Peyré, G., and Dossal, C. (2017). The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 69(4):791–832.

[215] Vaiter, S., Golbabaee, M., Fadili, J., and Peyré, G. (2015). Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287.

[216] Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

[217] Vohl, D., Fluke, C. J., and Vernardos, G. (2015). Data compression in the petascale astronomy era: A GERLUMPH case study. *Astronomy and Computing*, 12:200–211.

[218] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

[219] Wibisono, A., Wilson, A. C., and Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358.

[220] Wilson, A. C., Recht, B., and Jordan, M. I. (2016). A Lyapunov analysis of momentum methods in optimization. *arXiv e-prints*, page arXiv:1611.02635.

[221] Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 1(151):3–34.

[222] Wright, S. J. and Lee, C.-p. (2017). Analyzing random permutations for cyclic coordinate descent. *ArXiv e-prints*.

[223] Yin, W., Osher, S., Goldfarb, D., and Darbon, J. (2008). Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging sciences*, 1(1):143–168.

[224] Young, D. M. and Rheinboldt, W. (1971). *Iterative solution of large linear systems*. Academic Press.

[225] Zhang, X., Burger, M., and Osher, S. (2011). A unified primal-dual algorithm framework based on Bregman iteration. *Journal of Scientific Computing*, 46(1):20–46.

[226] Zhu, M. and Chan, T. (2008). An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 34.

Appendix A

Miscellaneous results

A.1 Cutoff function

We provide proof of existence of an appropriate cutoff function in Theorem 3.4 *(iii)*. While this is based on standard arguments using mollifiers, the authors could not find a result in the literature for cutoff functions with noncompact support and controlled derivatives. We therefore include one for completeness.

Lemma A.1. *Let $V, W \subset \mathbb{R}^n$ be disjoint (not necessarily compact) sets such that, for some $\varepsilon > 0$,*

$$\|x-y\| \geq \varepsilon, \quad \textit{for all } x \in V, y \in W.$$

Then there is a cutoff function $\varphi \in C^\infty(\mathbb{R}^n;[0,1])$ such that

$$\varphi(x) = \begin{cases} 1 & \textit{if } x \in V, \\ 0 & \textit{if } x \in W, \end{cases} \tag{A.1}$$

and such that $\nabla\varphi$ is uniformly bounded on $\mathbb{R}^n$.

Proof. We will construct a cutoff function with a uniformly bounded gradient. Consider the distance functions

$$d_V(x) := \inf_{z \in V} \|x-z\|, \qquad d_W(x) := \inf_{z \in W} \|x-z\|.$$

For any $x \in \mathbb{R}^n$, $y \in V$, $z \in W$, it holds that $\varepsilon \leq \|y-z\| \leq \|x-y\| + \|x-z\|$. Taking the infimum over all $y \in V$ and $z \in W$, we deduce that

$$d_V(x) + d_W(x) \geq \varepsilon. \tag{A.2}$$

Let $\psi : \mathbb{R}^n \to [0,1]$ be defined by $\psi(x) := d_W(x)/(d_V(x) + d_W(x))$. This function satisfies $\psi(X) = 1$ for $x \in V$, $V(x) = 0$ for $x \in W$ and $\psi(x) \in [0,1]$ otherwise. We will show that it is Lipschitz continuous with Lipschitz constant $1/\varepsilon$.

$$\begin{aligned}
|\psi(x) - \psi(y)| =& \left| \frac{d_W(x)}{d_V(x) + d_W(x)} - \frac{d_W(y)}{d_V(y) + d_W(y)} \right| \\
\leq& \frac{\big|d_V(y) - d_V(x)\big|\, d_W(x) + \big|d_W(x) - d_W(y)\big|\, d_V(x)}{\big(d_V(x) + d_W(x)\big)\, \big(d_V(y) + d_W(y)\big)} \\
\overset{(A.2)}{\leq}& \frac{1}{\varepsilon}\|x-y\| \left(\frac{d_W(x)}{d_V(x) + d_W(x)} + \frac{d_V(x)}{d_V(x) + d_W(x)} \right) = \frac{1}{\varepsilon}\|x-y\|.
\end{aligned}$$

The second inequality above follows from (A.2) and Lipschitz continuity of d_V and d_W.

We choose an appopriate mollifier $J \in C_c^\infty(\mathbb{R}^n; [0,\infty))$ such that $\int_{\mathbb{R}^n} J(x)\, \mathrm{d}x = 1$ and $J(x) \equiv 0$ outside $\overline{B}_{\varepsilon/2}(0)$, and convolve it with ψ. It is easy to check that the resultant function,

$$\varphi(x) = \int_{\mathbb{R}^n} J(z)\, \psi(x-z)\, \mathrm{d}z,$$

is in $C^\infty(\mathbb{R}^n; [0,1])$ and satisfies (A.1). This is a standard result, e.g. [4, Theorem 2.29]. To conclude, we show that $\|\nabla\varphi(x)\| \leq 1/\varepsilon$ for all $x \in \mathbb{R}^n$. We do so by showing that φ inherits the Lipschitz continuity of ψ. We have

$$\begin{aligned}
|\varphi(x) - \varphi(y)| &\leq \int_{\mathbb{R}^n} \big|\psi(x-z) - \psi(y-z)\big|\, |J(z)|\, \mathrm{d}z \\
&\leq \frac{1}{\varepsilon}\|x-y\| \int_{\mathbb{R}^n} |J(z)|\, \mathrm{d}z = \frac{1}{\varepsilon}\|x-y\|.
\end{aligned}$$

This concludes the proof. □

A.2 Convergence rate for cyclic coordinate descent

In what follows, we obtain improved convergence rates for cyclic coordinate descent (CCD) [15, 221] that match those obtained for the Itoh–Abe discrete gradient method in Section 3.6. The CCD method, for a starting point x^0, time steps $\tau_i > 0$, $i = 1, \ldots, n$, and $k = 0, 1, 2, \ldots$ is given by

$$\begin{aligned}
x^{k,0} &= x^k, \\
x^{k,i+1} &= x^{k,i} - \tau_{i+1} [\nabla F(x^{k,i})]_{i+1} e^{i+1}, \qquad \text{for } i = 0, \ldots, n-1, \\
x^{k+1} &= x^{k,n}.
\end{aligned} \tag{A.3}$$

Recalling Section 3.6, we are interested in estimates for $\beta > 0$ that satisfy (3.18), where smaller β implies better convergence rate. In [15] (see Lemma 3.3) and referenced in [221], the estimate

$$\beta = 4L_{\max}\left(1 + nL^2/L_{\min}^2\right),$$

is obtained, using the time step $\tau_i = 1/L_i$. This rate is optimised with respect to $L_{\min}, L_{\max}$ when setting $L_{\min} = L_{\max} = \sqrt{n}L$, yielding $\beta = 8\sqrt{n}L$. However, we show in Section 3.6.1 that the closely related Itoh–Abe discrete gradient method achieves the stronger bound $\beta = 4L_{\text{sum}} \leq 4\sqrt{n}L$. We therefore include a brief analysis to demonstrate that the bound for CCD can similarly be improved.

By the coordinate-wise descent lemma (3.19), we have

$$\begin{aligned} F(x^{k,i}) - F(x^{k,i+1}) &\geq \langle \nabla F(x^{k,i}), x^{k,i} - x^{k,i+1} \rangle - \frac{L_i}{2}\|x^{k,i} - x^{k,i+1}\|^2 \\ &= \left(\tau_i - \frac{\tau_i^2 L_i}{2}\right) |[\nabla F(x^{k,i})]_{i+1}|^2. \end{aligned}$$

For some $\alpha \in (0,2)$, we choose the time steps $\tau_i = \alpha/L_i$, and substitute into the above inequality to get

$$F(x^{k,i}) - F(x^{k,i+1}) \geq \frac{1}{L_i}\left(\alpha - \frac{\alpha^2}{2}\right) |[\nabla F(x^{k,i})]_i|^2. \tag{A.4}$$

We then compute

$$\begin{aligned}
\|\nabla F(x^k)\|^2 &= \sum_{i=1}^{n} |[\nabla F(x^k)]_i|^2 \\
&\leq 2\sum_{i=1}^{n} \Big(|[\nabla F(x^k)]_i - [\nabla F(x^{k,i-1})]_i|^2 + |[\nabla F(x^{k,i-1})]_i|^2 \Big) \\
&\overset{(A.4)}{\leq} 2\sum_{i=1}^{n} \left(L^2 \|x^k - x^{k,i}\|^2 + \frac{L_i}{\alpha - \frac{\alpha^2}{2}} \Big(F(x^{k,i-1}) - F(x^{k,i}) \Big) \right) \\
&\leq 2\sum_{i=1}^{n} \left(L^2 \sum_{j=0}^{i} \|x^{k,j} - x^{k,j+1}\|^2 + \frac{L_i}{\alpha - \frac{\alpha^2}{2}} \Big(F(x^{k,i-1}) - F(x^{k,i}) \Big) \right) \\
&\leq 2 \left(\frac{n\alpha^2 L^2}{L_{\min}^2} \sum_{j=0}^{n} |[\nabla F(x^{k,j})]_{j+1}|^2 + \frac{L_{\max}}{\alpha - \frac{\alpha^2}{2}} \Big(F(x^k) - F(x^{k+1}) \Big) \right) \\
&\leq \frac{2L_{\max}(1 + n\alpha^2 L^2 / L_{\min}^2)}{\alpha - \frac{\alpha^2}{2}} \Big(F(x^k) - F(x^{k+1}) \Big).
\end{aligned}$$

This gives a new estimate for β,

$$\beta = \frac{2L_{\max}(1 + n\alpha^2 L^2 / L_{\min}^2)}{\alpha - \frac{\alpha^2}{2}}.$$

If we set $\alpha = 1/\sqrt{n}$ and $L_i = L$, we get the estimate

$$\beta = 4L\sqrt{n} \left(\frac{2\sqrt{n}}{2\sqrt{n} - 1} \right) \approx 4\sqrt{n}L.$$

This is approximately the same rate as that obtained for the Itoh–Abe discrete gradient method.

It is too longwinded to compute the optimal values of τ_i and L_i to include it here, but one can confirm the optimal rate is close to the above estimate and satisfies

$$\frac{\beta^*}{\sqrt{n}} \to 4L \quad \text{as } n \to \infty.$$

Coordinate descent methods are typically extended to *block coordinate descent* methods. The above analysis can be extended to this setting simply by replacing n with the number of blocks p.