PARTLY SMOOTH MODELS AND ALGORITHMS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Calvin James Stuart Wylie

December 2019

PARTLY SMOOTH MODELS AND ALGORITHMS

Calvin James Stuart Wylie, Ph.D.

Cornell University 2019

Optimization and variational problems typically involve a highly structured blend of smooth and nonsmooth geometry. In nonlinear programming, such structure underlies the design of active-set algorithms, in which a globally convergent process first simplifies the problem by identifying active constraints at the solution; a second phase then employs a rapidly-convergent Newton-type method, with linear models of the simplified problem playing a central role. The theory of partial smoothness formalizes and highlights the fundamental geometry driving "identification." This dissertation concentrates on the second phase, and understanding accelerated local convergence in partly smooth settings.

A key contribution is a simple algorithm for "black-box" nonsmooth optimization, that incorporates second-order information with the usual linear approximation oracle. Motivated by active sets and sequential quadratic programming, a model-based approach is used to prove local quadratic convergence for a broad class of objectives. Promising numerical results on more general functions, as well as simple first-order analogues, are discussed. Beyond optimization, an intuitive linearization scheme for generalized equations is formalized, with simple techniques based on classical differential geometry: manifolds, normal and tangent spaces, and constant rank maps. The approach illuminates fundamental geometric ideas behind active-set acceleration techniques for variational inequalities, as well as second-order theory and algorithms for structured nonsmooth optimization.

BIOGRAPHICAL SKETCH

Calvin James Stuart Wylie was born in Ottawa, Canada, in 1989 and grew up in Canada and Northern Ireland. He attended the University of British Columbia where he received a Bachelor of Science in Mathematics in 2011. After three years as a software engineer, he began his doctoral studies in Operations Research at Cornell University in 2014. After graduation he will join Wayfair in Boston, Massachusetts, as a Senior Operations Research Scientist.

ACKNOWLEDGEMENTS

First and foremost, I thank my advisor Adrian Lewis for his guidance and support during the course of my studies. Without his mathematical knowledge, intuition, and creativity, this thesis would not have been possible. I also thank my committee members Damek Davis and David Bindel for helpful discussions, suggestions, and comments.

I thank the School of Operations Research for support over the years, as well as all the students, faculty, and staff who make ORIE the very best place to do a Ph.D.

For teaching me mathematics, introducing me to research, and encouraging me to pursue doctoral studies, I am deeply indebted to my professors Shawn Wang and especially Heinz Bauschke.

Last but certainly not least, I thank my parents Barbara and Peter for their never-ending love and encouragement, and Kelsey and Finn, whose support and love made this work possible.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The algorithmic idea of approximating the solution to a difficult problem with a sequence of solutions to easier problems has a long and rich history in mathematics. A particularly early example is the Babylonian method of computing square roots. More generally, we can often formulate a problem as finding a solution x to a system of equations in Euclidean space

$$F(x) = 0.$$

Works of Vieta, Newton, Raphson, and Simpson eventually gave rise to the first modern formulation – in the language of calculus – of what is now most commonly known as *Newton's method*. The algorithm is motivated by the fact that the derivative provides a good approximation to the equations of interest:

$$F(x + z) \approx F(x) + \nabla F(x)\, z \quad \text{for all } x \text{ and small } z.$$

Therefore, given a candidate solution x, we instead solve a system where F is replaced with the *local model* $F(x) + \nabla F(x)(\,\cdot - x)$ to find a new candidate solution. The resulting Newton iteration

$$x \leftarrow x - \nabla F(x)^{-1} F(x)$$

forms the core of countless numerical procedures due to its simplicity, intuitiveness, and good theoretical and practical performance.

However, situations abound in applied mathematics where the problem or equation system of interest is not only not differentiable but may lack any classical notion of continuity. Central to modern theory is the idea of *set-valued mappings*, with the problem of interest being a *generalized equation*

$$0 \in \Phi(x).$$

Such mappings play fundamental roles in the study of optimization and variational inequalities. Broadly, this dissertation is concerned with understanding local models, with the immediate aim of designing algorithms, in this setting. Mathematical analysis has historically been plagued by pathological counterexamples, and nonsmooth analysis is no exception. This has led researchers to seek suitable classes of structured problems, for example *convexity*, *semialgebraicity*, or *semismoothness*, that provide interesting and relevant analysis, while remaining general enough to encompass a broad landscape of practical applications. Of particular importance in this work is the structure of *partial smoothness*. This theory generalizes the notion of active sets in nonlinear programming to rigorously root the observation that sets and functions arising in practice are typically a highly structured blend of smooth and nonsmooth geometry.

One of the central contributions of this dissertation is providing a framework for understanding local linearization models and algorithms for generalized equations in a partly smooth setting. The techniques are fundamentally simple and based on classical differential geometry: manifolds, normal and tangent spaces, and constant rank maps. Despite this simplicity the framework illuminates a variety of interesting applications, particularly in optimization and variational inequalities.

Our second key contribution is to provide a practical and fully implementable Newton-type algorithm for local nonsmooth optimization, with the first known superlinear convergence result for a broad class of nonsmooth functions. Preliminary work and numerical experiments suggest several promising

future directions.

1.1 Smooth and nonsmooth models in optimization

A core numerical problem in applied mathematics and science is the accurate local minimization of a continuous real-valued objective function $f : \mathbf{R}^n \to \mathbf{R}$. Standard optimization literature [106, 9, 104, 31] initially assumes very little on the global structure of f: given a point $x \in \mathbf{R}^n$, we are able to compute the objective value $f(x)$, possibly the gradient $\nabla f(x)$, and possibly the Hessian $\nabla^2 f(x)$. When f is sufficiently smooth, a complete theory is well understood. Local optimization techniques rely on the gradient to define an accurate linear model $f(x) + \nabla f(x)^\mathsf{T}(\cdot - x)$ of the function around points of interest.

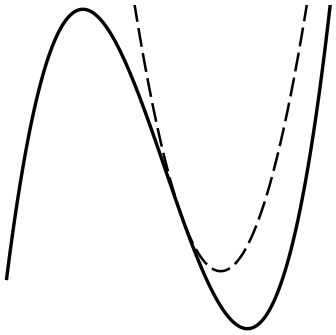


Figure 1.1: A quadratic model.

By minimizing the linear model augmented with a quadratic regularization term (suitably chosen to ensure steps that are neither too large or too small),

$$f(x) + \nabla f(x)^\mathsf{T}(\cdot - x) + \frac{\rho}{2}|\cdot - x|^2,$$

we recover *gradient descent*: we choose $x - \rho\nabla f(x)$ as a point with improved objective function value. Global convergence to stationary points can be achieved with appropriate line search or trust region techniques [106].

Assuming second-order smoothness allows us to derive conditions to determine which stationary points are minimizers. Furthermore, fast local rates of convergence can be achieved by applying Newton's method to the stationarity condition $\nabla f(x) = 0$. In a region around a local minimizer, the Newton iteration

$$x \leftarrow x - \nabla^2 f(x)^{-1} \nabla f(x)$$

is well-defined and has an alternative interpretation as the minimizer of the local *quadratic model*

$$f(x) + \nabla f(x)^\mathsf{T}(\cdot - x) + \frac{1}{2}(\cdot - x)^\mathsf{T} \nabla^2 f(x)(\cdot - x).$$

More complex *quasi-Newton* algorithms, which aim to strike a balance between the computational economy of gradient-based methods and the powerful convergence rates of Newton's method, are the traditional workhorses of numerical optimization code and remain an active area of research [82, 10, 34].

In the *nonsmooth* case, when the objective function is not differentiable everywhere (and in particular not necessarily at minimizers), the situation is more complicated. *Convex* sets and functions [115] have long been recognized as particularly amenable to computation and analysis, by appealing to the existence of separating hyperplanes and *subgradients*,

$$g \in \partial f(x) \;\Leftrightarrow\; f(z) \geq f(x) + g^\mathsf{T}(z - x) \;\text{ for all } z,$$

which yield one sided estimates of the objective function. Knowledge of the entire subdifferential $\partial f(x)$ is usually too stringent a requirement in practice, so algorithms for *unstructured* or "black box" convex optimization instead assume that a single subgradient can be computed via a subgradient oracle $g(x) \in \partial f(x)$. (The vast field of *structured* convex optimization, in which explicit representations of f are assumed, for example linear or conic programming, is beyond the scope of this work and not treated here.)

Cutting plane methods [27, 64] aim to construct a global model of the objective function by incorporating information from multiple subgradients. Specifically, given a finite set of points $X \subset \mathbf{R}^n$ and a subgradient oracle g, the cutting plane model of f is the piecewise linear function

$$y \mapsto \max_{x \in X} \{ f(x) + g(x)^T (y - x) \}.$$

The fundamental idea is that the cutting plane model becomes an increasingly accurate lower model as more points are added to the set X.

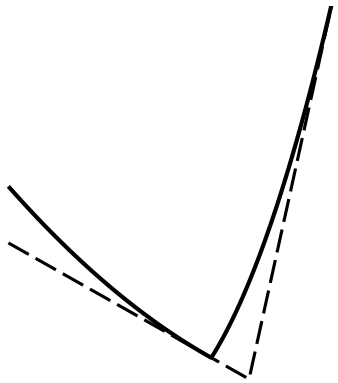


Figure 1.2: A cutting plane model.

Naive cutting plane implementations suffer from instability and poor practical performance [104, 5], but more sophisticated cutting plane techniques can be effective at moderate scale [2, 126, 76, 103, 74].

Proximal algorithms, based on the proximal mapping

$$x \mapsto \arg\min_y \{ f(y) + \frac{1}{2} |y - x|^2 \}$$

introduced in the seminal works of Moreau [100, 101] and popularized by the proximal point algorithm for monotone operations of Rockafellar [116], play fundamental roles in modern optimization theory and applications [111, 30,

11, 83]. For convex optimization, proximal algorithms and their various operator splitting generalizations [3] possess more favourable convergence properties than subgradient-based algorithms [5]. However, computation of the proximal point is itself a nonsmooth optimization problem, and is thus not necessarily easier than the original problem. Successful application is therefore limited to cases where the objective function can be decomposed into smooth and "prox-friendly" ingredients. In other words, nonsmoothness is handled analytically and not modeled numerically.

Proximal cutting plane or *bundle methods* [75, 129, 93, 68] (see also the survey [107]) are a combination of cutting plane and proximity control ideas, and involve subproblem models of the form

$$y \mapsto \max_{x \in X} \left\{ f(x) + g(x)^{\mathsf{T}}(y - x) \right\} + \frac{\rho}{2} |y - z|^2$$

for a *bundle* of points X and subgradients $g(x) \in \partial f(x)$ for $x \in X$. With appropriate parameter selection and judicious updates of the *center* z, a bundle method can be viewed as an "implementable proximal point algorithm" [120]. Despite their popularity over the past several decades, results on convergence rates for bundle methods are sparse [67, 44] and several aspects of the behaviour of the method remain poorly understood [96].

In the nonsmooth and nonconvex setting, designing even local minimization algorithms is much more difficult. Bundle methods have been extended to the nonconvex case [54, 69, 92, 122, 128], however implementation is delicate, and nonconvexity is handled in a heuristic way that is not supported by strong convergence theory or well-defined models of the objective function. A recent line of work [58, 57] improves upon previous nonconvex bundle methods by considering the class of *prox-regular* functions (nonconvex functions that admit well-

defined proximal points), but no convergence rates are given.

When the objective function is locally Lipschitz, an alternate direction is to appeal to Rademacher's theorem and work with the *Clarke subdifferential* [29]

$$\partial_c f(\bar{x}) = \operatorname{conv}\left\{\lim_{r\to\infty} \nabla f(x_r) : x_r \to \bar{x}, x_r \in \mathcal{D}\right\},$$

where $\mathcal{D}$ is the (full measure) set of points where f is differentiable. Stationary points can thus be recognized when zero is a convex combination of gradients of nearby points. Gradient sampling [18] strategies attempt to model the Clarke subdifferential as

$$\partial_c f(\bar{x}) \approx \operatorname{conv}\{\nabla f(x) : x \text{ randomly sampled near } \bar{x}\}.$$

With large enough samples, this approximation can then be used to generate descent directions, leading to algorithms that are globally convergent to stationary points with high probability.

1.2 Partial smoothness

Even in the convex case, nonsmooth algorithms based on subgradients are theoretically limited to poor rates of convergence [104]. Despite this, the folklore in the optimization community (verified experimentally during the course of writing this dissertation) is that bundle methods usually perform much better in practice than the theory suggests. One reason for this may be that typical nonsmooth functions are highly structured; minimizers tend to lie on "ridges" of the nonsmooth graph. This, in fact, holds generically for semialgebraic functions [40].

The theory of *partial smoothness* [79] formalizes this with the existence of an *active manifold* $\mathcal{M}$ relative to which the restricted function $f\big|_{\mathcal{M}}$ is smooth. Or-

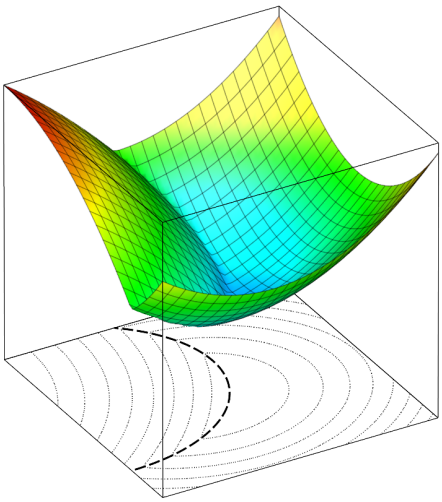


Figure 1.3: A partly smooth function.

thogonal to $\mathcal{M}$ the function behaves in a nonsmooth manner. A simple example is the two-dimensional function

$$f(u, v) = u^2 + |v| \qquad (u, v \in \mathbf{R}),$$

which relative to the manifold $\mathbf{R} \times \{0\}$ behaves as the smooth univariate function $t \mapsto t^2$. In the language of $\mathcal{VU}$-theory [77, 97, 94], we decompose $\mathbf{R}^2$ into the complementary $\mathcal{U}$-subspace $\mathbf{R} \times \{0\}$ and $\mathcal{V}$-subspace $\{0\} \times \mathbf{R}$.

Based on these decompositions, one can envision a conceptual algorithm that learns $\mathcal{M}$ [98, 84] so as to employ a smooth model on the reduced function $f\big|_{\mathcal{M}}$. An implementable algorithm combining the equivalent (in the convex case) $\mathcal{VU}$-theory with a typical bundle method is given in [95], but the method has several drawbacks. Implementation is complicated, and since $\mathcal{VU}$-theory and bundle methods are only well understood for convex functions, the algorithm does not immediately generalize to nonconvex objectives. Fast convergence is also only proven for a sequence of "serious steps," and a bound on the total computation required to reach accurate solutions remains unknown.

In contrast to these carefully structured approaches for fast nonsmooth minimization, another line of work [82] investigates the baffling success of applying quasi-Newton algorithms to nonsmooth functions. A rigorous theory does not exist, but there is a large amount of numerical evidence, that the BFGS algorithm automatically identifies partly smooth structure. When applied to nonsmooth functions, the Hessian approximations generated by the algorithm become increasingly ill-conditioned precisely in directions orthogonal to the active manifold ("$\mathcal{V}$-space"). Some recent algorithmic frameworks [32, 34] have found practical success in employing gradient sampling techniques to improve the accuracy of solutions found via BFGS.

In Chapter 3 we take a fresh look at partly smooth models for nonsmooth optimization with a novel semi-structured approach, driven by considering the simple but broad class of *max functions* of the form

$$f(x) = \max_i f_i(x)$$

that are partly smooth with respect to the manifold

$$\mathcal{M} = \{x : f_i(x) \text{ equal for all } i\}.$$

We show how it is easy to estimate the dimension of $\mathcal{M}$ using existing algorithms, and armed only with this dimension k, we develop a new class of algorithms that converge at local quadratic rates by incorporating second-order information, and without full knowledge of the underlying structure functions f_i. We also discuss some preliminary first-order extensions of this model.

1.3 Beyond optimization: Generalized equations

Generalizing Newton's method beyond smooth equations has long been an active area of research. We refer the reader to the monographs [49, 38, 61], or the recent survey [62]. In the case of a nonsmooth equation $F(x) = 0$, one avenue is to work with the *Clarke generalized Jacobian*

$$\partial F(\bar{x}) = \operatorname{conv}\left\{ \lim_{r\to\infty} \nabla F(x_r) : x_r \to \bar{x} \right\}.$$

For the class of *semismooth* functions (see e.g., [49]), elements of the Clarke Jacobian define an adequate Newton model in the sense that

$$\lim_{\substack{x\to\bar{x} \\ G\in\partial F(x)}} \frac{F(x) + G(\bar{x} - x) - F(\bar{x})}{|x - \bar{x}|} = 0,$$

and therefore we can consider Newton iterations of the form

$$x \leftarrow x - G^{-1}F(x) \quad \text{for } G \in \partial F(x).$$

Since semismooth functions arise broadly and naturally [7], semismooth Newton methods [113] have enjoyed broad practical success, especially in the infinite-dimensional setting [125].

In the set-valued or generalized equation setting, equations often take the form $0 \in F(x) + \Psi(x)$ where F is smooth and Ψ is set-valued. In this setting the Josephy-Newton method [63] can be employed which considers the partially linearized model equation

$$0 \in F(x) + \nabla F(x)(\,\cdot - x) + \Psi(\,\cdot\,).$$

(Obvious extensions to semismooth F can also be considered.) Under suitable "metric regularity" assumptions – which generalize nonsingularity of the Jacobian in classical Newton's method – these methods are well-defined and fast

rates of convergence can be established [38, 62]. However, a limitation of these methods is that they ultimately rely on linearizing single-valued mappings, and leave the set-valued ingredient alone, precluding their application to generalized equations unless the problem can be reformulated so that Ψ is simple enough to work with directly. As an example, consider the optimization problem of minimizing a linear function $x \mapsto c^{\mathsf{T}}x$ over the convex constraint set

$$K = \{x : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m\}.$$

Minimizers $\bar{x}$ satisfy the generalized stationarity equation $-c \in N_K(\bar{x})$. The *normal cone* operator N_K is a complicated object, but by introducing dual variables $y \in \mathbf{R}^m$ and a *Lagrangian*, assuming a constraint qualification we can write the stationarity condition as

$$0 \in \begin{pmatrix} c + \nabla g(x)y \\ g(x) \end{pmatrix} + N_E(x, y),$$

where E is the simpler set $\mathbf{R}^n \times \mathbf{R}^m_+$. Applying the Josephy-Newton model captures the class *sequential quadratic programming* methods. But this generality fails to capture the practically important *active-set* philosophy: when the active constraints $A(\bar{x})$ (those j for which $g_j(\bar{x}) = 0$) have been identified, the problem reduces to minimization over the smooth lower-dimensional constraint set

$$\{x : g_j(x) = 0 \text{ for } j \in A(\bar{x})\},$$

for which direct Newton methods can be employed.

In Chapter 4, we consider generalized equations in a partly smooth setting. Based on the recent extension [80] of partial smoothness to set-valued maps, we develop an intuitive linearization scheme in broad generality. The approach illuminates fundamental geometric ideas behind active-set algorithms for varia-

tional inequalities and higher-order schemes for modern composite optimization.

We end this dissertation in Chapter 5 with a return to unstructured nonsmooth minimization. Based on insights gained from the previous chapter, we modify the algorithms of Chapter 3 to derive an implementable partly smooth Newton algorithm. While a rigorous convergence theory is left as a topic for future research, promising numerical results are shown.

CHAPTER 2

PRELIMINARIES

This chapter lays out notation and collects preliminary definitions and well-known results that will play important roles in this dissertation.

2.1 Euclidean space

Our setting is that of finite-dimensional inner product (or Euclidean) spaces, denoted **E**, **U**, etc., over the real numbers **R**. We denote the inner product $\langle \cdot\,, \cdot \rangle$ and corresponding induced norm $|\cdot| = \sqrt{\langle \cdot\,, \cdot \rangle}$. The unit ball is fixed as

$$B = \{x : |x| \leq 1\}$$

and the ball of radius $\delta > 0$ centered at z as

$$B_\delta(z) = \{x : |x - z| \leq \delta\}.$$

Inner products (and corresponding norms) of Cartesian product spaces $\mathbf{U} \times \mathbf{V}$ are defined in the natural way as

$$\langle (u, v), (w, z) \rangle = \langle u, w \rangle + \langle v, z \rangle.$$

We define the operator norm on the space of (necessarily bounded) linear maps $T : \mathbf{U} \to \mathbf{V}$ as

$$|T| = \sup_{u \in \mathbf{U}} \frac{|Tu|}{|u|}.$$

We say a set $\{x_0, x_1, \ldots, x_n\} \subset E$ is *linearly independent* if

$$\sum_{i=0}^{n} \lambda_i x_i = 0 \implies \lambda = 0$$

and *affinely independent* if $\{x_1 - x_0, \ldots, x_n - x_0\}$ is linearly independent. Given a linear map $A : \mathbf{U} \to \mathbf{V}$ the *adjoint* is the unique linear map $A^* : \mathbf{V} \to \mathbf{U}$ such that

$$\langle Au, v\rangle = \langle u, A^*v\rangle.$$

A self-adjoint ($A = A^*$) map is called *positive semidefinite* if $\langle u, Au\rangle \geq 0$ for all $u \in \mathbf{U}$. If the inequality is strict for all $u \neq 0$ we say A is *positive definite*.

We write $O(t)$ to denote a term satisfying

$$\limsup_{t\to 0} \frac{O(t)}{t} < \infty$$

and $o(t)$ to denote a term satisfying

$$\lim_{t\to 0} \frac{o(t)}{t} = 0.$$

If z is a vector, we sometimes write $O(z)$ and $o(z)$ to mean $O(|z|)$ and $o(|z|)$ respectively.

Given a sequence x_k converging to $\bar{x}$, we say that the convergence rate is *linear* if there exists some $r \in (0, 1)$ such that

$$|x_{k+1} - \bar{x}| \leq r\,|x_k - \bar{x}| \quad \text{for large } k,$$

superlinear if

$$|x_{k+1} - \bar{x}| = o\big(\,|x_k - \bar{x}|\big),$$

and *quadratic* if

$$|x_{k+1} - \bar{x}| = O\big(\,|x_k - \bar{x}|^2\big).$$

2.2 Smooth maps

Let $F : \mathbf{U} \to \mathbf{V}$ be a map between Euclidean spaces $\mathbf{U}$ and $\mathbf{V}$. We say F is *differentiable* at $u \in \mathbf{U}$ if there exists a linear map $T : \mathbf{U} \to \mathbf{V}$ such that

$$F(u + z) = F(u) + Tz + o(z).$$

When F is differentiable at u we call T the *derivative* of F at u, and denote it $DF(u)$. When the map $u \mapsto DF(u)$ is continuous and defined for all $u \in \mathbf{U}$ we say that F is *continuously differentiable*, $C^{(1)}$-smooth, or just $C^{(1)}$. Since DF maps $\mathbf{U}$ to linear operators $\mathbf{U} \to \mathbf{V}$, we can define higher order derivatives and smoothness in an identical manner.

For $F : \mathbf{R}^n \to \mathbf{R}^m : x \mapsto (f_1(x), \ldots, f_m(x))$, the derivative $DF(x)$ can be represented as the *Jacobian matrix*

$$\begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_m} \end{pmatrix}$$

$C^{(k)}$-smoothness then amounts to the existence and continuity of all kth order partial derivatives.

For real valued $f : \mathbf{E} \to \mathbf{R}$, the derivative takes the form $Df(x) : z \mapsto \langle g, z \rangle$ for some $g \in \mathbf{E}$ which we call the *gradient* $\nabla f(x)$. $C^{(2)}$-smoothness amounts to the functions $h_z : x \mapsto \langle \nabla f(x), z \rangle$ being $C^{(1)}$-smooth for all z. In this case we call the bilinear operator $\nabla^2 f(x)[w, z] = \langle w, \nabla h_z(x) \rangle$ the *Hessian* of f.

When $\mathbf{E} = \mathbf{R}^n$ with the usual dot product $\langle u, v \rangle = u^\mathsf{T} v$ the gradient can be represented as the column vector of partial derivatives $\left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$ and the Hessian takes the form $\nabla^2 f(x)[w, z] = w^\mathsf{T} \nabla^2 f(x) z$ for the *Hessian matrix*

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_m^2} \end{pmatrix}.$$

The following results are central and will be used throughout, their proofs can be found in any advanced calculus text.

Theorem 2.2.1 (Taylor's Theorem). *For twice continuously differentiable $f : \mathbf{E} \to \mathbf{R}$,*

$$f(x + tz) = f(x) + t \langle \nabla f(x), z \rangle + \frac{t^2}{2} \nabla^2 f(x)[z, z] + o(t^2)$$

for all $x, z \in \mathbf{E}$.

Proposition 2.2.1. *For $f : \mathbf{E} \to \mathbf{R}$ with ∇f L-Lipschitz,*

$$\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| \leq \frac{L}{2} |y - x|^2$$

for all $x, y \in \mathbf{E}$.

Corollary 2.2.1. *For twice continuously differentiable $F : \mathbf{E} \to \mathbf{R}^m$,*

$$F(u + z) = F(u) + DF(u)z + O(|z|^2).$$

2.3 Convex and nonsmooth analysis

Following standard variational analysis [119], we denote the extended real line $\bar{\mathbf{R}} = \mathbf{R} \cup \{+\infty, -\infty\}$, with the convention that $(+\infty) + (-\infty) = +\infty$ and $0 \times (\pm\infty) = 0$. We denote the *domain* of an extended valued function $f : \mathbf{E} \to \bar{\mathbf{R}}$ by

$$\operatorname{dom} f = \left\{ x \in \mathbf{E} : f(x) < +\infty \right\}.$$

f is *proper* if it never takes the value $-\infty$ and $\operatorname{dom} f$ is nonempty. f is *closed* or *lower semicontinuous* if its *epigraph*

$$\operatorname{epi} f = \left\{ (x, t) \in \mathbf{E} \times \mathbf{R} : t \geq f(x) \right\}$$

is a closed set. The *indicator* function of a set $S \subset \mathbf{E}$ is defined by

$$\delta_S(x) = \begin{cases} 0 & (x \in S), \\ +\infty & (x \notin S). \end{cases}$$

A point $y \in \mathbf{E}$ is a *regular subgradient* of f at $x \in \operatorname{dom} f$ if

$$f(x+z) \geq f(x) + \langle y, z \rangle + o(z).$$

The set of all regular subgradients at x is denoted $\hat{\partial} f(x)$. y is a *(limiting) subgradient* if there exist sequences (x_r) and (y_r) with $y_r \in \hat{\partial} f(x_r)$ such that $x_r \to x$, $f(x_r) \to f(x)$, and $y_r \to y$. We say that f is *regular* at $x \in \operatorname{dom} f$ if $\partial f(x) \neq \varnothing$ and $\hat{\partial} f(x) = \partial f(x)$. If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

For a closed set $S \subset \mathbf{E}$, the *regular normal cone* to S at x is

$$\hat{N}_S(x) = \hat{\partial} \delta_S(x) = \left\{ y \in \mathbf{E} : \langle y, z - x \rangle \leq o(|z - x|) \text{ for all } z \in S \right\}.$$

The *(limiting) normal cone* is $N_S(x) = \partial \delta_S(x)$ and S is regular at x when $N_S(x) = \hat{N}_S(x)$. The *tangent cone* to S at x is

$$T_S(x) = \left\{ \lim_{r \to \infty} d_r : \exists t_r \searrow 0 \text{ with } x + t_r d_r \in S \right\}.$$

A set $S \subset \mathbf{E}$ is *convex* if $\lambda x + (1-\lambda) y \in S$ for all $x, y \in S$ and $0 \leq \lambda \leq 1$. S is *affine* if this holds for any real λ. The *convex hull* $\operatorname{conv} S$ is the smallest convex set containing S. The affine hull is defined identically and the *relative interior* $\operatorname{ri} S$ is the interior of S with respect to its affine hull.

A function $f : \mathbf{E} \to \bar{\mathbf{R}}$ is convex if $\operatorname{epi} f$ is a convex set. f is *μ-strongly convex* if $f - \frac{\mu}{2} |\cdot|^2$ is convex, and *η-weakly convex* if $f + \frac{\eta}{2} |\cdot|^2$ is convex. Convex sets and functions are everywhere regular in the sense that

$$N_S(x) = \hat{N}_S(x) = \{ y : \langle y, z - x \rangle \leq 0 \text{ for all } z \in S \},$$
$$\partial f(x) = \hat{\partial} f(x) = \{ y : f(x) + \langle y, z - x \rangle \leq f(z) \text{ for all } z \in \mathbf{E} \}.$$

Theorem 2.3.1. *A closed convex and proper function $f : \mathbf{E} \to \mathbf{R}$ is locally Lipschitz around any $x \in \operatorname{int}(\operatorname{dom} f)$, and moreover $\varnothing \neq \partial f(x) \subset LB$ where L is a local Lipschitz constant.*

Theorem 2.3.2 ([29], Proposition 2.2.7 and Theorem 2.5.1). *When $f : \mathbf{E} \to \mathbf{R}$ is locally Lipschitz and regular at $\bar{x}$,*

$$\partial f(\bar{x}) = \operatorname{conv}\left\{\lim_{r\to\infty} \nabla f(x_r) : x_r \to \bar{x}, x_r \in \mathcal{D}, x_r \notin \mathcal{S} \text{ for } r = 1, 2, \ldots\right\}$$

where $\mathcal{S} \subset \mathbf{E}$ is any set of Lebesgue measure zero.

Normal cones and subdifferentials motivate the analysis of *set-valued* mappings. The most convenient way to work with a set-valued mapping $\Phi : \mathbf{U} \rightrightarrows \mathbf{V}$ is through its *graph*

$$\operatorname{gph}\Phi = \{(u, v) \in \mathbf{U} \times \mathbf{V} : v \in \Phi(u)\}.$$

Definition 2.3.1. The *graphical derivative* of Φ at u for v, $D\Phi(u, v) : \mathbf{U} \rightrightarrows \mathbf{V}$, is characterized by

$$z \in D\Phi(u|v)(w) \Leftrightarrow (w, z) \in T_{\operatorname{gph}\Phi}(u, v).$$

The *coderivative* of Φ at u for v, $D^*\Phi(u, v) : \mathbf{V} \rightrightarrows \mathbf{U}$, is characterized by

$$w \in D^*\Phi(u|v)(z) \Leftrightarrow (w, -z) \in N_{\operatorname{gph}\Phi}(u, v).$$

Graphical derivatives and coderivatives can be viewed as a rudimentary kind of generalized differentiation, with DF and D^*F coinciding with the classical derivative and its adjoint when $F : \mathbf{U} \to \mathbf{V}$ is a smooth singled-valued mapping, and satisfying the following basic calculus.

Theorem 2.3.3 ([119], see 10.43). *Consider set-valued $\Psi : \mathbf{U} \rightrightarrows \mathbf{V}$ and $C^{(1)}$-smooth $F : \mathbf{U} \to \mathbf{V}$. Then for any u and v,*

$$D(F + \Psi)(u|v)(w) = \nabla F(u)w + D\Psi(u|v - F(u))(w) \quad \text{for all } w,$$
$$D^*(F + \Psi)(u|v)(y) = \nabla F(u)^* y + D^*\Psi(u|v - F(u))(y) \quad \text{for all } y.$$

2.4 Smooth manifolds

Loosely speaking, a manifold is a set that is locally homeomorphic to Euclidean space. A little more precisely, $\mathcal{X}$ is a manifold around z if there is some homeomorphism between $\mathcal{X}$ and a *coordinate space*, $\phi : \mathcal{X} \to \mathbf{R}^k$, locally defined around z, such that $\phi(z) = 0$. We call ϕ a *coordinate map* and say it is *centered* around z, and we say that the *dimension* of $\mathcal{X}$ is k. The following definitions and results are standard [72].

The most useful manifolds for the purposes of this work are those that are the solution sets of smooth equations with linearly independent gradients. Specifically, we will be concerned with $C^{(r)}$*-smooth embedded submanifolds* (for some $r = 1, 2, \ldots$), and let the term $C^{(r)}$-manifold refer to those of this type.

Definition 2.4.1. $\mathcal{X}$ is a $C^{(r)}$-manifold around $z \in \mathbf{E}$ if and only if there exists a $C^{(r)}$-smooth *defining map* $G : \mathbf{E} \to \mathbf{R}^m$ with $DG(z)$ surjective such that $G^{-1}(0)$ is a neighbourhood of z in $\mathcal{X}$.

A manifold defined by $G : \mathbf{E} \to \mathbf{R}^m$ has dimension $\dim \mathbf{E} - m$ and *codimension* m. It will sometimes be useful to work with an equivalent definition that makes use of coordinate space.

Definition 2.4.2. $\mathcal{X}$ is a $C^{(r)}$-manifold around $z \in \mathbf{E}$ if and only if there exists a $C^{(r)}$-smooth *local parametrization* $H : \mathbf{R}^k \to \mathbf{E}$ such that $H(0) = z$, with $DH(0)$ injective and $H(\delta B)$ a neighbourhood of z in $\mathcal{X}$ for all small $\delta > 0$.

The tangent and normal cones of a smooth manifold coincide with the classical tangent and normal *spaces*, which have the following algebraic representa-

tions for defining map G and local parametrization $H(w) = x$.

$$T_{\mathcal{X}}(x) = \text{Range}(DH(w)) = \text{Null}(DG(x)),$$
$$N_{\mathcal{X}}(x) = \text{Null}(DH(w)^*) = \text{Range}(DG(x)^*).$$

It is important to note that $T_{\mathcal{X}}(x)$ and $N_{\mathcal{X}}(x)$ are geometric objects independent of any specific choice of G and H.

Proposition 2.4.1. *Let $\mathcal{X}$ be a $C^{(1)}$-manifold around $z \in \mathbf{E}$ with local parametrization $H : \mathbf{R}^k \to \mathbf{E}$. If $u, v \in \mathbf{R}^k$ are sufficiently small then*

$$u - v = O\big(H(u) - H(v)\big) \quad \textit{and} \quad H(u) - H(v) = O(u - v).$$

Proof. Since $DH(0)$ is injective, H is locally a bijection between $\mathbf{R}^k$ and $\mathcal{X}$, so there exists some smooth inverse $H^{-1} : \mathcal{X} \to \mathbf{R}^k$ such that $H^{-1}(H(w)) = w$ for all small w. Thus,

$$\begin{aligned} |u - v| &\leq \big|H^{-1}\big|\big|H(u) - H(v)\big| \\ &\leq \big|H^{-1}\big|\big|H\big|\,|u - v|. \end{aligned}$$

The result follows immediately. □

Definition 2.4.3. Given a smooth map $P : \mathcal{X} \to \mathcal{Y}$ between manifolds $\mathcal{X}, \mathcal{Y} \subset \mathbf{E}$, the *rank* of P at $\bar{u}$ is defined to be the rank of the derivative

$$DP(\bar{u}) : T_{\mathcal{X}}(\bar{u}) \to T_{\mathcal{Y}}(F(\bar{u})).$$

If in some neighbourhood of $\bar{u}$ the rank of P is constant, we say that P is constant rank near $\bar{u}$.

The constant rank theorem is fundamental. It says that constant rank maps look like simple projection maps when viewed in appropriate coordinate spaces.

Theorem 2.4.1 (Constant Rank Theorem). *Let $\mathcal{X}$ and $\mathcal{Y}$ be smooth manifolds of dimension m, n respectively, and $F : \mathcal{X} \to \mathcal{Y}$ a smooth map with constant rank k. Then for each $x \in \mathcal{X}$ there exists local parametrizations $\phi : \mathbf{R}^k \times \mathbf{R}^{m-k} \to \mathcal{X}$ centered at x and $\psi : \mathbf{R}^k \times \mathbf{R}^{n-k} \to \mathcal{Y}$ centered at $F(x)$ such that*

$$F(\phi(w, u)) = \psi(w, 0) \quad \textit{for small } w, u.$$

2.5 Partial smoothness

The definition of partial smoothness first appeared in [79]. Here we adopt the slightly modified definitions appearing in [42, 80]. We must first impose the regularity conditions of *prox-regularity*, which bridges convex sets and smooth manifolds, and *subdifferential continuity* ([119, see 13.F]). In the convex case, these conditions hold automatically.

Definition 2.5.1. A closed function $f : \mathbf{E} \to \bar{\mathbf{R}}$ is *prox-regular* at $\bar{u}$ for value $\bar{v} \in \partial f(\bar{u})$ if $f(\bar{u})$ is finite, and there exists a $\rho \geq 0$ and $\epsilon > 0$ such that

$$f(u') \geq f(u) + \langle v, u' - u \rangle - \frac{\rho}{2} |u' - u|^2 \quad \text{for all } u' \in B_\epsilon(\bar{u})$$

when $u \in B_\epsilon(\bar{u})$, $v \in \partial f(u) \cap B_\epsilon(\bar{v})$, and $f(u) < f(\bar{u}) + \epsilon$.

Definition 2.5.2. A function $f : \mathbf{E} \to \bar{\mathbf{R}}$ is *subdifferentially continuous* at $\bar{u}$ for $\bar{v} \in \partial f(\bar{u})$ if any sequence $(u_k, v_k) \to (\bar{u}, \bar{v})$ with $v_k \in \partial f(u_k)$ also satisfies $f(u_k) \to f(\bar{u})$.

Definition 2.5.3. A closed set $Q \subset \mathbf{E}$ is *prox-regular* at $\bar{u} \in Q$ for value $\bar{v} \in N_Q(\bar{u})$ when δ_Q is prox-regular at $\bar{u}$ for $\bar{v}$, i.e., there exists a $\rho \geq 0$ and $\epsilon > 0$ such that

$$\langle v, u' - u \rangle \leq \frac{\rho}{2} |u' - u|^2 \quad \text{for all } u' \in Q \cap B_\epsilon(\bar{u})$$

when $u \in B_\epsilon(\bar{u})$ and $v \in N_Q(u) \cap B_\epsilon(\bar{v})$.

In particular, prox-regularity of Q at $\bar{u}$ for 0 implies that the projection mapping Proj_Q is single-valued around $\bar{u}$.

Definition 2.5.4. Consider a closed function $f : \mathbf{E} \to \bar{\mathbf{R}}$, a $C^{(r)}$-manifold $\mathcal{M}$ around $\bar{u} \in \mathbf{E}$ and a subgradient $\bar{v} \in \partial f(\bar{u})$. We that that f is *$C^{(r)}$-partly smooth at $\bar{u}$ for $\bar{v}$ relative to $\mathcal{M}$* if

- f is prox-regular at $\bar{u}$ for $\bar{v}$.
- f is $C^{(r)}$-smooth around $\bar{u}$ relative to $\mathcal{M}$.
- $\operatorname{span} \hat{\partial} f(\bar{u}) = N_{\mathcal{M}}(\bar{u}) + \bar{v}$.
- For any $v \in \partial f(\bar{u})$ near $\bar{v}$, and any sequence $u_k \in \mathcal{M}$ converging to $\bar{u}$, there exists a sequence $v_k \in \partial f(u_k)$ converging to v.

Definition 2.5.5. Consider a closed set $Q \subset \mathbf{E}$, a $C^{(r)}$-manifold $\mathcal{M}$ around $\bar{u} \in \mathbf{E}$ and a normal vector $\bar{v} \in N_Q(\bar{u})$. We say that Q is *$C^{(r)}$-partly smooth at $\bar{u}$ for $\bar{v}$ relative to $\mathcal{M}$* if the following holds.

- Q is prox-regular at $\bar{u}$ for $\bar{v}$.
- $\operatorname{span} \hat{N}_Q(\bar{u}) = N_{\mathcal{M}}(\bar{u})$.
- For any $v \in N_Q(\bar{u})$ near $\bar{v}$, and any sequence $u_k \in \mathcal{M}$ converging to $\bar{u}$, there exists a sequence $v_k \in N_Q(u_k)$ converging to v.

In particular, Q is partly smooth if δ_Q is partly smooth.

Definition 2.5.6. A set-valued mapping $\Phi : \mathbf{U} \rightrightarrows \mathbf{V}$ is *$C^{(r)}$-partly smooth at $\bar{u}$ for $\bar{v} \in \Phi(\bar{u})$* if $\operatorname{gph} \Phi$ is a $C^{(r)}$-manifold around $(\bar{u}, \bar{v})$, and the projection

$$P : \operatorname{gph} \Phi \to \mathbf{U} : (u, v) \mapsto u$$

is constant rank in a neighbourhood W of $(\bar{u}, \bar{v})$. We call

$$\mathcal{M} = P(\operatorname{gph}\Phi \cap W)$$

the *active manifold*.

Partly smooth set-valued mappings are related to partly smooth sets and functions in the following way.

Theorem 2.5.1 ([80], Theorems 5.3 and 5.5). *The following are equivalent for a closed set $Q \subset \mathbf{E}$ and corresponding normal cone mapping N_Q.*

(i) *Q is $C^{(r)}$-partly smooth at $\bar{u}$ for $\bar{v}$ relative to $\mathcal{M}$, and $\bar{v} \in \operatorname{ri} N_Q(\bar{u})$.*

(ii) *N_Q is $C^{(r-1)}$-partly smooth at $\bar{u}$ for $\bar{v}$ with active manifold $\mathcal{M}$.*

(iii) *$\operatorname{gph} N_Q = \operatorname{gph} N_{\mathcal{M}}$ in a neighbourhood of $(\bar{u}, \bar{v})$.*

The following are equivalent for a closed function $f : \mathbf{E} \to \bar{\mathbf{R}}$ and corresponding subdifferential mapping ∂f.

(i) *f is subdifferentially continuous at $\bar{u}$ for $\bar{v}$, $C^{(r)}$-partly smooth at $\bar{u}$ for $\bar{v}$ relative to $\mathcal{M}$, and $\bar{v} \in \operatorname{ri} \partial f(\bar{u})$.*

(ii) *∂f is $C^{(r-1)}$-partly smooth at $\bar{u}$ for $\bar{v}$ with active manifold $\mathcal{M}$.*

(iii) *In a neighbourhood of $(\bar{u}, \bar{v})$,*

$$\operatorname{gph} \partial f = \{(u, \nabla \bar{f}(u) + v) : u \in \mathcal{M}, v \in N_{\mathcal{M}}(u)\},$$

where $\bar{f}$ is any $C^{(r)}$-smooth function agreeing with f on $\mathcal{M}$.

Example (Partly smooth function and subdifferential). Consider the proper closed convex function

$$f(x, y) = x^2 + |y|,$$

and its subdifferential

$$\partial f(x,y) = \left\{ (u,v) : u = 2x, v = \begin{cases} 1 & \text{if } y > 0 \\ [-1,1] & \text{if } y = 0 \\ -1 & \text{if } y < 0 \end{cases} \right\},$$

which satisfies $(0,0) \in \operatorname{ri} \partial f(0,0)$. Now, f is convex, therefore everywhere prox-regular, and on the manifold

$$\mathcal{M} = \{(x,y) : y = 0\}$$

it agrees with the smooth function $(x,y) \mapsto x^2$. Also, $\operatorname{span} \partial f(0,0) = \{(u,v) : u = 0\}$, which is the normal space $N_{\mathcal{M}}(0,0)$. Inner semicontinuity relative to $\mathcal{M}$ clearly holds, so f is partly smooth at $(0,0)$ for $(0,0)$.

Locally, $\operatorname{gph} \partial f$ around $((0,0),(0,0))$ is parametrized by

$$(w,z) \mapsto (w,0,2w,z),$$

which is a linear subspace, hence a manifold around 0. The projection onto the first two coordinates is the subspace parametrized by $w \mapsto (w,0)$, which has constant dimension 1, so ∂f is partly smooth at $(0,0)$ for $(0,0)$ with active manifold $\mathcal{M}$.

Example (Partly smooth set and normal cone). Suppose that Q is closed and convex (therefore everywhere prox-regular) with the representation

$$Q = \{u \in \mathbf{U} : g_i(u) \leq 0 \text{ for } i = 1, \ldots, n\},$$

where the functions g_i are smooth. Denote the *active constraints*

$$A(\bar{u}) = \{i : g_i(\bar{u}) = 0\},$$

and assume the *constraint qualification* that the active gradients $\{\nabla g_i(\bar{u})\}_{i\in A(\bar{u})}$ are linearly independent. Then

$$N_Q(\bar{u}) = \left\{ \sum_{i\in A(\bar{u})} \lambda i \nabla g_i(\bar{u}) : \lambda \geq 0 \right\},$$

and defining the manifold

$$\mathcal{M} = \{u \in \mathbf{U} : g_i(u) = 0 \text{ for } i \in A(\bar{u})\},$$

we have that $N_{\mathcal{M}}(\bar{u}) = \operatorname{span} N_Q(\bar{u})$. Given a sequence $u_k \in \mathcal{M}$ converging to $\bar{u}$, and a normal vector $\bar{v} = \sum_{i\in A(\bar{u})} \bar{\lambda}_i \nabla g_i(\bar{u})$, observe that we can construct a corresponding sequence of normal vectors $v_k \in N_Q(u_k)$ by setting $v_k = \sum_{i\in A(\bar{u})} \bar{\lambda}_i \nabla g_i(u_k)$, which converge to $\bar{v}$ by continuity. Hence Q is partly smooth.

The condition $\bar{v} \in \operatorname{ri} N_Q(\bar{u})$ becomes the *strict complementarity* condition that the multipliers $\bar{\lambda}_i$ of the active constraints corresponding to the normal vector $\bar{v}$ are uniformly bounded away from zero, which implies that N_Q is partly smooth with $P(\operatorname{gph} N_Q) = \mathcal{M}$ around $(\bar{u}, \bar{v})$.

CHAPTER 3

BUNDLE NEWTON ALGORITHMS

3.1 Introduction

In this chapter, we develop a local algorithm for finding a minimizer of a continuous nonsmooth objective function. We assume that the objective $f : \mathbf{E} \to \mathbf{R}$ is smooth around every point in some set $\mathcal{D} \subset \mathbf{E}$, and that at any point $x \in \mathcal{D}$ we can access an *oracle* that returns the value $f(x)$, gradient $\nabla f(x)$, and Hessian $\nabla^2 f(x)$.

First-order algorithms in this "black-box" model typically fall in one of three categories: subgradient, cutting plane, or random sampling based. Subgradient methods (originating in [124]), comprised of iterations of the form

$$x \leftarrow x - \frac{\rho}{|\nabla f(x)|} \nabla f(x) \qquad (x \in \mathcal{D}),$$

are perhaps the most simplistic class of algorithms in this setting, with broad general purpose appeal. In the nonsmooth setting, the gradient does not necessarily vanish near a minimizer, so the step sizes ρ are typically chosen in advance and scaled towards zero. This leads a notoriously slow convergence rate [102, 104, 36], both in theory and practice, limiting the basic subgradient iteration to applications where accuracy is not a major concern, extremely large scale settings where the computational burden of more complex methods is too great, or problems with benign structure [37].

For convex functions, variations of bundle methods [129, 75] and level-set bundle methods [76], which reuse previous gradient information, are often the method of choice when gradient computations are expensive and solution accu-

racy is a concern [107]. In theory, bundle methods converge at sublinear rates [67, 44], and generate difficult to analyze sequences of "null steps." Though fast convergence on a sequence of "serious steps" is possible [95], this drawback of bundle methods remains an obstacle.

Gradient sampling [18, 33, 12] can be effective when gradients are relatively cheap to compute, and readily extends to nonconvex settings, but does not scale well to high dimensions. Based on the practical success of quasi-Newton methods applied to nonsmooth objectives [82], some recent work [59, 34] combines gradient sampling with quasi-Newton ideas to develop practical and robust minimization routines. However, no convergence rates are known for these methods.

Still missing in nonsmooth optimization is the simplicity and fast local convergence of Newton's method. The aim of this chapter is a step in this direction: a simple black-box local optimization method, supported by rigorous theory, and able to incorporate second-order information to achieve a superlinear rate of convergence. We take a semi-structured approach, and first consider minimizing a pointwise maximum of finitely many smooth functions, using a black-box oracle that cannot access the underlying component functions individually. (Such a model was also analyzed in [60] in the context of gradient sampling algorithms.) Our resulting algorithm converges at a local quadratic rate on nonsmooth functions of this type, and in Chapter 5, we show promising experimental results on more general functions. We also give some preliminary first-order extensions of the algorithm. The majority of the chapter appears in the manuscript [87].

To keep the motivation and development clean, at the outset we focus on a strongly convex objective. Extending to a broad class of nonconvex objectives

turns out to be relatively straightforward, and is handled in following sections. Specifically, we are able to handle weakly convex functions in a manner similar to the bundle method [57].

3.2 Convex minimization

Techniques for smooth minimization rely critically on Fermat's rule that $\nabla f(\bar{x}) = 0$ at a minimizer, and the principle that $|\nabla f(x)|$ is small when x is close to $\bar{x}$. In the nonsmooth setting, we cannot in general hope to find a point with a small gradient, and the generalized stationarity condition $0 \in \partial f(\bar{x})$ is far too stringent to verify in our black box setting. Instead, we will use local information at multiple points to build an approximate optimality measure. To borrow the terminology of bundle methods, we seek a finite *bundle* $S \subset \mathcal{D}$ of *reference points* with small *diameter*

$$\operatorname{diam} S = \max_{s,s' \in S} |s - s'| \tag{3.2.1}$$

and small *optimality measure*

$$\Theta(S) = \min \left| \operatorname{conv} \left(\nabla f(S) \right) \right|. \tag{3.2.2}$$

By defining the unit simplex

$$\Delta_S = \left\{ \lambda \in \mathbf{R}_+^{|S|} : \sum_{s \in S} \lambda_s = 1 \right\},$$

we can equivalently write the optimality measure as

$$\Theta(S) = \min_{\lambda \in \Delta_S} \left| \sum_{s \in S} \lambda_s \nabla f(s) \right|. \tag{3.2.3}$$

Below we describe an algorithm that iteratively updates a bundle one reference point at a time. The minimizing λ of the optimality measure $\Theta(S)$, which

we will later interpret as a Lagrange multiplier, plays an important role in defining the core subproblem of the algorithm: minimizing a weighted average of local quadratic approximations

$$q_s(\cdot) = f(s) + \langle \nabla f(s), \cdot - s \rangle + \frac{1}{2} \langle \cdot - s, \nabla^2 f(s)(\cdot - s) \rangle$$

on a subspace defined using the local linear approximations

$$l_s(\cdot) = f(s) + \langle \nabla f(s), \cdot - s \rangle.$$

Algorithm 3.1: Local Newton algorithm to minimize convex f

Input : Bundle $S \subset \mathcal{D}$, tolerances $\bar{\epsilon}, \bar{\delta} \geq 0$;
while diam $S > \bar{\delta}$ and $\Theta(S) > \bar{\epsilon}$ **do**
  for $s \in S$ **do**
    $l_s(\cdot) = f(s) + \langle \nabla f(s), \cdot - s \rangle$;
    $q_s(\cdot) = l_s(\cdot) + \frac{1}{2} \langle \cdot - s, \nabla^2 f(s)(\cdot - s) \rangle$;
  end
  Choose $\lambda \in \Delta_S$ to minimize $\left| \sum_{s \in S} \lambda_s \nabla f(s) \right|$;
  Choose $\hat{x} \in \arg\min \left\{ \sum_{s \in S} \lambda_s q_s(x) : l_s(x) \text{ equal for all } s \in S \right\}$;
  if $x \notin \mathcal{D}$ **then**
    Stop;
  else
    Choose $s \in S$ minimizing $\Theta\big((S \setminus s) \cup \hat{x}\big)$;
    $S \leftarrow (S \setminus s) \cup \hat{x}$;
  end
end

In contrast to most cutting plane and bundle methods, a crucial feature is the fixed size of the bundle S, which is given as input to the algorithm. With an input of $|S| = 1$ the method becomes exactly classical Newton's method for smooth minimization. The bundle size must be judiciously chosen. With too small a choice the algorithm may fail to converge to a minimizer, as does Newton's method applied to the nonsmooth objective $\max((x-1)^2, (x+1)^2)$, which cycles between non-optimal points $\{\pm 1\}$. A bundle size chosen too large causes a type of degeneracy and algorithmic instability that we discuss in later sections.

3.2.1 The optimality measures

The algorithm aims to construct a sequence of bundles S converging to a minimizer $\bar{x}$ (and hence $\operatorname{diam} S$ converging to zero), such that the optimality measure $\Theta(S)$ also converges to zero. In practice, we can terminate when these measures are small to deduce approximate optimality in the following way.

Given a bundle S let λ be the minimizing multiplier associated with optimality measure (3.2.3), and define the weighted average of the reference points

$$\bar{s} = \sum_{s \in S} \lambda_s s.$$

By convexity, the linear approximations minorize f:

$$l_s(x) \;=\; f(s) + \langle \nabla f(s), x - s \rangle \;\leq\; f(x) \quad \text{for all } x \in \mathbf{E},$$

and being continuous and convex, f is also locally Lipschitz. Let us denote by L a Lipschitz constant for f on some open ball containing S, so that we may bound the size of the gradients:

$$|\nabla f(s)|^2 \;=\; \langle \nabla f(s), s - (s - \nabla f(s)) \rangle \;\leq\; f(s - \nabla f(s)) - f(s) \;\leq\; L\,|\nabla f(s)|,$$

which implies $|\nabla f(s)| \leq L$. Also, by convexity of the norm,

$$\begin{aligned}
\sum_{s \in S} \lambda_s \,|s - \bar{s}| &\leq \max_{s \in S} |s - \bar{s}| \\
&\leq \max_{s \in S} \sum_{s' \in S} \lambda_{s'} \,|s - s'| \\
&\leq \max_{s \in S} \max_{s' \in S} |s - s'| \\
&= \operatorname{diam} S.
\end{aligned}$$

Altogether, we have that for all $x \in \mathbf{E}$

$$\begin{aligned} f(\bar{s}) - L\,\mathrm{diam}(S) &\leq f(\bar{s}) - L\sum_{s\in S}\lambda_s\,|s-\bar{s}| \\ &\leq f(\bar{s}) + \sum_{s\in S}\lambda_s\,\langle \nabla f(s), \bar{s}-s\rangle \\ &\leq \sum_{s\in S}\lambda_s l_s(\bar{s}) \\ &= \sum_{s\in S}\lambda_s l_s(x) + \sum_{s\in S}\langle \lambda_s \nabla f(s), \bar{s}-x\rangle \\ &\leq \sum_{s\in S}\lambda_s l_s(x) + \Theta(S)\,|x-\bar{s}| \\ &\leq f(x) + \Theta(S)\,|x-\bar{s}|, \end{aligned}$$

which implies

$$\min f \;\leq\; f(\bar{s}) \;\leq\; \min\left\{f + \Theta(S)\,|\cdot - \bar{s}|\right\} + L\,\mathrm{diam}(S). \tag{3.2.4}$$

In other words, if the diameter and optimality measures are small, then the current bundle constitutes an approximate certificate of optimality in the sense that the point $\bar{s}$ lies between $\min f$ and the minimum of a slightly perturbed function.

While this bound motivates the optimality measures, it is less useful computationally. If a strong convexity constant is known, it is possible to terminate with a guaranteed optimality gap. In addition to the trivial upper bound

$$\min f \;\leq\; \bar{f} \;=\; \min_{s\in S} f(s),$$

if f is ρ-strongly convex we also have the lower bound

$$\min f \;\geq\; \underline{f} \;=\; \min_{x\in\mathbf{E}}\left\{\sum_{s\in S}\lambda_s\left(l_s(x) + \frac{\rho}{2}\,|x-s|^2\right)\right\},$$

which is easy to compute since the right hand side is minimized at

$$\underline{x} = \bar{s} - \frac{1}{\rho}\sum_{s\in S}\lambda_s \nabla f(s).$$

When the gap $\bar{f} - \underline{f}$ is less than some tolerance $\tau > 0$, it holds that

$$\min_{s \in S} f(s) < \min f + \tau.$$

To see that this condition will hold whenever $\operatorname{diam} S$ and $\Theta(S)$ are sufficiently small, observe that for all $s \in S$,

$$\begin{aligned} l_s(\underline{x}) &= l_s(\bar{s}) - \frac{1}{\rho} \left\langle \nabla f(s), \sum_{s' \in S} \lambda_{s'} \nabla f(s') \right\rangle \\ &\geq l_s(\bar{s}) - \frac{L}{\rho} \Theta(S) \\ &\geq l_s(\bar{s}) - L \left| \bar{s} - s \right| - \frac{L}{\rho} \Theta(S) \\ &\geq f(s) - L \operatorname{diam} S - \frac{L}{\rho} \Theta(S) \\ &\geq \bar{f} - L \operatorname{diam} S - \frac{L}{\rho} \Theta(S). \end{aligned}$$

Hence we deduce that

$$\begin{aligned} \bar{f} - \underline{f} &= \bar{f} - \sum_{s \in S} \lambda_s \left(l_s(\underline{x}) + \frac{\rho}{2} \left| \underline{x} - s \right|^2 \right) \\ &\leq \bar{f} - \sum_{s \in S} \lambda_s l_s(\underline{x}) \\ &\leq L \operatorname{diam} S + \frac{L}{\rho} \Theta(S). \end{aligned} \tag{3.2.5}$$

3.2.2 A lower bound on bundle size

For the algorithm to succeed, the optimality measure $\Theta(S)$ must converge to zero for a sequence of bundles converging to minimizer $\bar{x}$. In other words, zero must be a convex combination of the limiting gradient set

$$\Gamma = \left\{ \lim_{r \to \infty} \nabla f(x_r) : x_r \to \bar{x}, x_r \in \mathcal{D} \text{ for } r = 1, 2, \ldots \right\}. \tag{3.2.6}$$

A lower bound on the bundle size required for the algorithm to succeed is thus the minimum size of a subset of Γ whose convex hull contains zero: the

Carathéodory number $\operatorname{car}\Gamma$, which by Carathéodory's theorem satisfies

$$1 \;\leq\; \operatorname{car}\Gamma \;\leq\; \dim(\operatorname{conv}\Gamma).$$

(Note that since $\mathcal{D}$ is full measure when f is continuous and convex, the convex hull of Γ is simply the subdifferential $\partial f(\bar{x})$.)

3.2.3 The active subspace

Equipped with $\Theta(S)$ and associated multipliers λ, the algorithm next seeks a new reference point. A standard cutting plane approach would consider the model $\tilde{f} : \mathbf{E} \to \mathbf{R}$ defined by

$$\tilde{f}(x) = \max_{s \in S} l_s(x),$$

which minorizes f and approximates it up to first order at each of the reference points. At every point on the *active subspace*

$$M = \{x \in \mathbf{E} : l_s(x) \text{ equal for all } s \in S\},$$

the cutting plane model has subdifferential

$$\partial \tilde{f}(x) = \operatorname{conv}(\nabla f(S)),$$

and hence nonsmooth slope equal to $\Theta(S)$. Therefore when the optimality measure is small, the cutting plane model is approximately minimized throughout M, so this is where we seek a new reference point.

3.2.4 An upper bound on bundle size

Since the new reference point lies in the active subspace M, a basic requirement for algorithmic stability is that $\bar{x}$ is close to M when $\max |S - \bar{x}|$ (and hence

in particular diam S) is small. By convexity

$$0 \;\le\; f(\bar{x}) - l_s(\bar{x}) \;=\; f(\bar{x}) - f(s) + \langle \nabla f(s), s - x \rangle \;\le\; 2L\,|s - \bar{x}| \quad \text{for all } s \in S,$$

so $f(\bar{x})$ is close to $l_s(\bar{x})$. Therefore we equivalently ask that the point $(\bar{x}, f(\bar{x})) \in \mathbf{E} \times \mathbf{R}$ be close to the affine subspace

$$\{(x,t) \in \mathbf{E} \times \mathbf{R} : l_s(x) = t \text{ for all } s \in S\}.$$

As we have observed, the residuals of the linear system defining this subspace are small at $(\bar{x}, f(\bar{x}))$ in the sense that

$$f(\bar{x}) - l_s(\bar{x}) = O(|s - \bar{x}|) \quad \text{for all } s \in S.$$

Standard linear algebra shows that this point is close to the nonempty solution set providing that the linear operator $L : (x,t) \mapsto (l_s(x) - t)_{s \in S}$ is full rank, which amounts to affine independence of the gradients $\nabla f(S)$.

Fixing coordinates $\mathbf{E} = \mathbf{R}^n$, uniform affine independence amounts to the matrix with columns $\binom{\nabla f(s)}{1}$ for $s \in S$ having smallest singular value greater than some tolerance $\sigma > 0$ for all bundles near $\bar{x}$. Success of the algorithm then requires an upper bound on the bundle size, since by taking a convergent subsequence of the matrices above we arrive at a limiting matrix with $|S|$ linearly independent columns of the form $\binom{g}{1}$, where each vector g is a limiting gradient and hence lies in the subdifferential $\partial f(\bar{x})$. Therefore the function f has at least $|S|$ affinely independent subgradients at $\bar{x}$, from which we deduce an upper bound of $1 + \dim(\partial f(\bar{x}))$ on the bundle size.

We will later prove that when the gradients $\nabla f(S)$ are affinely independent, the minimization problem (3.2.3) has a unique solution.

3.2.5 Choosing the bundle size

To summarize, if the algorithm succeeds, the bundle size must satisfy the bounds involving the Carathéodory number of the limiting gradient set (3.2.6) and the dimension of the subdifferential:

$$\operatorname{car}\Gamma \;\leq\; |S| \;\leq\; \dim(\partial f(\bar{x})) + 1.$$

In general these bounds may be far apart, but in some cases they are equal.

Example (Euclidean norm). Consider the Euclidean norm $\|\cdot\|_2$ and $\bar{x} = 0$. This function is smooth on $\mathcal{D} = \mathbf{R}^n \setminus \{0\}$ with gradient $\nabla \|\cdot\|_2(s) = \frac{1}{\|s\|_2} s$ for $s \neq 0$, so the limiting gradient set is simply the unit sphere, and subdifferential $\partial \|\cdot\|_2(\bar{x})$ the closed unit ball. Hence we have the bounds

$$2 \;\leq\; |S| \;\leq\; n + 1.$$

Example (Convex max functions). Consider a nonsmooth function of the form

$$f(x) = \max_{i=1,\ldots,k} f_i(x)$$

for smooth convex functions $f_i : \mathbf{E} \to \mathbf{R}$ for $i = 1, \ldots, k$, and a point $\bar{x} \in \mathbf{E}$ with function values $f_i(\bar{x})$ all equal, so that we have

$$\partial f(\bar{x}) = \operatorname{conv}\{\nabla f_i(\bar{x}) : i = 1, \ldots, k\}.$$

Assuming affine independence of the gradients $\{\nabla f_i\}$, we have

$$\dim(\partial f(\bar{x})) = k - 1.$$

Furthermore, assuming that $\bar{x}$ is a minimizer, so $0 \in \partial f(\bar{x})$, the system

$$\sum_{i=1}^{k} \lambda_i \nabla f_i(\bar{x}) = 0, \quad \sum_{i=1}^{k} \lambda_i = 1, \quad \lambda \in \mathbf{R}^k_+$$

must have a unique solution $\bar{\lambda}$. Since the limiting gradient set is

$$\Gamma = \{\nabla f_i(\bar{x}) : i = 1, \ldots, k\},$$

car Γ is simply the number of nonzero components of $\hat{\lambda}$, which is exactly k when $\bar{x}$ is a *nondegenerate* minimizer, meaning $0 \in \mathrm{ri}(\partial f(\bar{x}))$.

Estimating the lower bound on bundle size, car Γ, seems challenging in general. On the other hand, global nonsmooth optimization methods – such bundle methods, gradient sampling, and nonsmooth BFGS – typically suggest subdifferential dimension information. We defer a more detailed explanation to a future section, but the general idea is that given any finite set of points $\Omega \subset \mathcal{D}$ near the minimizer $\bar{x}$, motivated by the Clarke characterization of the subdifferential (Theorem 2.3.2) we can write the approximation

$$\mathrm{conv}(\nabla f(\Omega)) \approx \partial f(\bar{x}).$$

This suggests that a reasonable estimate of the dimension of $\partial f(\bar{x})$ is the numerical rank r of the matrix with columns $\binom{\nabla f(x)}{1}$ for $x \in \Omega$. A selection of r robustly affinely independent vectors in $\{\nabla f(x) : x \in \Omega\}$ might then serve as an initial bundle to our Newton method.

3.2.6 The quadratic subproblem

At the end of each iteration we update the bundle by swapping out some $s \in S$ for a new reference point $\hat{x}$ so as to minimize the optimality measure. The Newtonian flavour of the algorithm arises from how we choose $\hat{x}$, which solves a simple linearly-constrained quadratic program, and which we motivate with a model-based approach.

Since we assume the gradients $\nabla f(S)$ are affinely independent, this gradient information is inconsistent with any smooth model of f. We instead seek a simple nonsmooth model. Motivated by the previous example, we use a max function for the model, and consider smooth functions $f_s : \mathbf{E} \to \mathbf{R}$ each satisfying

$$f_s(s) = f(s), \quad \nabla f_s(s) = \nabla f(s), \quad \nabla^2 f_s(s) = \nabla^2 f(s), \quad f_s(s') < f(s')$$

for distinct $s, s' \in S$. Such functions always exist in theory, e.g. by defining

$$f_s(x) = q_s(x) - \alpha \left| x - s \right|^4 \qquad (x \in \mathbf{E})$$

for large enough $\alpha > 0$. In practice, their precise form is immaterial to the algorithm. We now consider the function $\tilde{f} : \mathbf{E} \to \mathbf{R}$ defined by

$$\tilde{f}(x) = \max_{s \in S} f_s(x)$$

as an (unknown) model of the objective function f, which agrees with f up to second order at each reference point. Minimizing this model is equivalent to solving the nonlinear program

$$\begin{aligned} \text{minimize} \quad & t \\ \text{subject to} \quad & f_s(x) - t \leq 0 \quad (s \in S) \\ & x \in \mathbf{E}, \quad t \in \mathbf{R}, \end{aligned}$$

which cannot be solved exactly since the functions f_s are unknown. Instead, we loosely follow a classical sequential quadratic programming approach to solve an approximation of the model, based on our reference bundle S.

One standard approach proceeds in two steps. The first would estimate the optimal Lagrange multipliers $\mu \in \Delta_S$ by minimizing $\left| \sum_{s \in S} \mu_s \nabla f_s(\tilde{x}) \right|$ at a trial

point $\tilde{x}$. We instead use our bundle of reference points to arrive at exactly the computation of the optimality measure

$$\Theta(S) = \min_{\lambda \in \Delta_S} \left| \sum_{s \in S} \lambda_s \nabla f(s) \right|.$$

Fixing the resulting Lagrange multiplier λ, the second step would then minimize a quadratic model of the Lagrangian

$$(x, t) \mapsto \sum_{s \in S} \lambda_s f_s(x)$$

over a feasible region defined by linearized constraints. Following again our philosophy of using information from multiple reference points, and restricting our attention to the active subspace as discussed in Section 3.2.3, we arrive at

$$\begin{aligned} \text{minimize} \quad & \sum_{s \in S} \lambda_s q_s(x) \\ \text{subject to} \quad & l_s(x) - t = 0 \quad (s \in S) \\ & x \in \mathbf{E}, \quad t \in \mathbf{R}, \end{aligned} \tag{3.2.7}$$

exactly the quadratic subproblem in the algorithm. This subproblem is feasible, as we saw in Section 3.2.4, and bounded below by our assumption of convexity.

3.3 Smooth-nonsmooth sums

In the previous section we restricted our attention to a Newton algorithm for convex functions to avoid complicating the motivation. The algorithm that we will analyze, however, is a version for minimization problems with composite objectives $F : \mathbf{E} \to \mathbf{R}$ of the form

$$F(x) = f(x) + r(x) \tag{3.3.1}$$

where $f : \mathbf{E} \to \mathbf{R}$ is nonsmooth but strongly convex, and $r : \mathbf{E} \to \mathbf{R}$ is $C^{(2)}$-smooth but possibly nonconvex. As previously, we assume that the nonsmooth

function f is twice continuously differentiable around every point in some set $\mathcal{D} \subset \mathbf{E}$.

The algorithm outlined below has one subtle change from the convex version Algorithm 3.1. While the quadratic approximations q_s and optimality measure

$$\Theta(S) = \min \left| \operatorname{conv} \left(\nabla F(S) \right) \right|$$

remain with respect to the overall objective, the linear approximations l_s that define the subproblem constraints are only with respect to the nonsmooth portion. In this nonconvex setting, we must also consider the possibility of the subproblem being unbounded, since the quadratic model is no longer necessarily convex. In this case, we halt the algorithm, since the bundle is most likely not near a local minimizer of F.

Algorithm 3.2: Local Newton algorithm to minimize composite F

Input : Bundle $S \subset \mathcal{D}$, tolerances $\bar{\epsilon}, \bar{\delta} \geq 0$;
while diam $S > \bar{\delta}$ and $\Theta(S) > \bar{\epsilon}$ **do**
  for $s \in S$ **do**
    $l_s(\cdot) = f(s) + \langle \nabla f(s), \cdot - s \rangle$;
    $q_s(\cdot) = F(s) + \langle \nabla F(s), \cdot - s \rangle + \frac{1}{2} \langle \cdot - s, \nabla^2 F(s)(\cdot - s) \rangle$;
  end
  Choose $\lambda \in \Delta_S$ to minimize $\left| \sum_{s \in S} \lambda_s \nabla F(s) \right|$;
  if $\min \left\{ \sum_{s \in S} \lambda_s q_s(x) : l_s(x) \textit{ equal for all } s \in S \right\} = -\infty$ **then**
    Stop;
  else
    Choose optimal $\hat{x}$;
    if $x \notin \mathcal{D}$ **then**
      Stop;
    else
      Choose $s \in S$ minimizing $\Theta\big((S \setminus s) \cup \hat{x}\big)$;
      $S \leftarrow (S \setminus s) \cup \hat{x}$;
    end
  end
end

The motivation remains largely the same, with a few slight changes. Since the objective function F is no longer convex, we of course cannot deduce a global optimality bound similar to (3.2.4). Instead we must settle for approximate Clarke stationarity.

To motivate the active subspace and quadratic program, we instead consider the partial cutting plane model

$$\tilde{F}(x) = \max_{s \in S} l_s(x) + r(x),$$

and the nonlinear program in Section 3.2.6 becomes

$$\begin{aligned} \text{minimize} \quad & t + r(x) \\ \text{subject to} \quad & f_s(x) - t \leq 0 \quad (s \in S) \\ & x \in \mathbf{E}, \quad t \in \mathbf{R}. \end{aligned}$$

We note that the constraints have not changed, so neither does the active subspace.

The discussion on bundle size applies in this new setting as well, by noting the simple relationship between the subdifferentials

$$\partial F(x) = \partial f(x) + \nabla r(x) \quad \text{for all } x \in \mathbf{E}.$$

3.4 A sequential quadratic programming tool

In this section we develop a tool which is relatively independent of the rest of the chapter, but will be critical in the analysis of our Newton algorithm. It is a slight variant of a classical sequential quadratic programming result for nonconvex nonlinear programming. For completeness, and because a suitable reference with our exact technical requirements could not be found, we prove it directly.

Given twice continuously differentiable objective function $f : \mathbf{E} \to \mathbf{R}$ and twice continuously differentiable constraints $g_i : \mathbf{E} \to \mathbf{R}$ for $i = 1, \ldots, k$, we consider the optimization problem

$$\begin{aligned} \text{minimize} \quad & f(y) \\ \text{subject to} \quad & g_i(y) = 0 \quad (i = 1, \ldots, k) \\ & y \in \mathbf{E}. \end{aligned} \tag{NLP}$$

Suppose that at a feasible point $\bar{y} \in \mathbf{E}$ the *constraint qualification*

$$G = \{\nabla g_i(\bar{y}) : i = 1, \ldots, k\} \text{ is linearly independent,}$$

stationarity condition that there exists a (necessarily unique) multiplier vector $\bar{\lambda} \in \mathbf{R}^k$ such that

$$\nabla f(\bar{y}) + \sum_{i=1}^{k} \bar{\lambda}_i \nabla g_i(\bar{y}) = 0,$$

and *second-order sufficient condition* that

$$\nabla^2 f(\bar{y}) + \sum_{i=1}^{k} \bar{\lambda}_i \nabla^2 g_i(\bar{y}) \text{ is positive definite on } G^{\perp}$$

hold. Fix any $\alpha \in \mathbf{R}^k$ such that $\sum_{i=1}^{k} \alpha_i = 1$. Given a collection of reference points $y_i \in \mathbf{E}$ and a multiplier estimate $\lambda \in \mathbf{R}^k$, define the quadratic program

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{k} \alpha_i \left[\langle \nabla f(y_i), y - y_i \rangle + \frac{1}{2} \langle y - y_i, \nabla^2 f(y_i)(y - y_i) \rangle \right] \\ & + \frac{1}{2} \sum_{i=1}^{k} \lambda_i \langle y - y_i, \nabla^2 g_i(y_i)(y - y_i) \rangle \\ \text{subject to} \quad & g_i(y_i) + \langle \nabla g_i(y_i), y - y_i \rangle = 0 \quad (i = 1, \ldots, k) \\ & y \in \mathbf{E}. \end{aligned} \tag{QP}$$

Theorem 3.4.1. *For any $\boldsymbol{y} \in \mathbf{E}^k$ near $\bar{\boldsymbol{y}} = (\bar{y}, \ldots, \bar{y})$ and any multiplier vector $\lambda = \bar{\lambda} + O(|\boldsymbol{y} - \bar{\boldsymbol{y}}|)$, the problem (QP) has a unique stationary point satisfying $\hat{y} = \bar{y} + O(|\boldsymbol{y} - \bar{\boldsymbol{y}}|^2)$, which furthermore is the unique minimizer.*

Proof. By continuity, the gradients $\{\nabla g_i(y_i) : i = 1, \ldots, k\}$ are also linearly independent, so necessary conditions for y to be a stationary point of (QP) are

$$\sum_{i=1}^{k} \alpha_i \left[\nabla f(y_i) + \nabla^2 f(y_i)(y - y_i)\right] + \sum_{i=1}^{k} \lambda_i \nabla^2 g_i(y_i)(y - y_i) = -\sum_{i=1}^{k} \mu_i \nabla g_i(y_i)$$

$$g_i(y_i) + \langle \nabla g_i(y_i), y - y_i \rangle = 0 \quad (i = 1, \ldots, k),$$

for some multiplier vector $\mu \in \mathbf{R}^k$. We can write these as a linear system

$$(M(\boldsymbol{y}, \lambda))(y, \mu) = b(\boldsymbol{y}, \lambda)$$

for linear operator $M(\boldsymbol{y}, \lambda) : \mathbf{E} \times \mathbf{R}^k \to \mathbf{E} \times \mathbf{R}^k$ and vector $b(\boldsymbol{y}, \lambda) \in \mathbf{E} \times \mathbf{R}^k$ depending continuously on parameter $(\boldsymbol{y}, \lambda)$. We claim that $M(\bar{\boldsymbol{y}}, \bar{\lambda})$ is invertible. Indeed, any solution to the homogeneous system $(M(\bar{\boldsymbol{y}}, \bar{\lambda}))(y, \mu) = 0$ satisfies

$$\left(\nabla^2 f(\bar{y}) + \sum_{i=1}^{k} \bar{\lambda}_i \nabla^2 g_i(\bar{y})\right) y + \sum_{i=1}^{k} \mu_i \nabla g_i(\bar{y}) = 0$$

$$\langle \nabla g_i(\bar{y}), y \rangle = 0 \quad (i = 1, \ldots k),$$

which implies that

$$\langle y, (\nabla^2 f(\bar{y}) + \sum_{i=1}^{k} \bar{\lambda}_i \nabla^2 g_i(\bar{y}))y \rangle = -\langle y, \sum_{i=1}^{k} \mu_i \nabla g_i(\bar{y}) \rangle = 0.$$

By the second-order sufficient conditions, $y = 0$, and hence $\mu = 0$ by linear independence of the constraint gradients.

As $\delta = |\boldsymbol{y} - \bar{\boldsymbol{y}}| \to 0$ with $\lambda - \bar{\lambda} = O(\delta)$, we have that for each $i = 1, \ldots, k$

$$g_i(y_i) + \langle \nabla g_i(y_i), \bar{y} - y_i \rangle \;=\; g_i(\bar{y}) + O(\delta^2) \;=\; O(\delta^2),$$

and also

$$\begin{aligned}
&\sum_{i=1}^{k} \alpha_i \left[\nabla f(y_i) + \nabla^2 f(y_i)(\bar{y} - y_i)\right] + \sum_{i=1}^{k} \lambda_i \nabla^2 g_i(y_i)(\bar{y} - y_i) \\
= \;& \nabla f(\bar{y}) + \sum_{i=1}^{k} \lambda_i \nabla g_i(\bar{y}) - \sum_{i=1}^{k} \lambda_i \nabla g_i(y_i) + O(\delta^2) \\
= \;& \sum_{i=1}^{k} (\lambda_i - \bar{\lambda}_i) \nabla g_i(\bar{y}) - \sum_{i=1}^{k} \lambda_i \nabla g_i(y_i) + O(\delta^2) \\
= \;& \sum_{i=1}^{k} -\bar{\lambda}_i \nabla g_i(y_i) + O(\delta^2).
\end{aligned}$$

Therefore

$$(M(\boldsymbol{y}, \lambda))(\bar{y}, \bar{\lambda}) - b(\boldsymbol{y}, \lambda) = O(\delta^2),$$

and because the norm of $M(\boldsymbol{y}, \lambda)$ is uniformly bounded for $(\boldsymbol{y}, \lambda)$ near $(\bar{\boldsymbol{y}}, \bar{\lambda})$,

$$(\bar{y}, \bar{\lambda}) - (M(\boldsymbol{y}, \lambda))^{-1} b(\boldsymbol{y}, \lambda) = O(\delta^2).$$

So there exists a unique stationary point $\hat{y} = \bar{y} + O(|\boldsymbol{y} - \bar{\boldsymbol{y}}|^2)$. Since we assume the second-order sufficient conditions, $\hat{y}$ is in fact the unique global minimizer of the equality constrained QP [106, Theorem 16.2]. □

In analyzing first-order variants of the algorithm, it will also be useful to refer to the following approximation to (NLP) and make use of a known result.

$$\begin{aligned}
\text{minimize} \quad & \langle \nabla f(z), y - z \rangle + \frac{1}{2} \langle y - z, B(y - z) \rangle \\
\text{subject to} \quad & g_i(z) + \langle \nabla g_i(z), y - z \rangle = 0 \quad (i = 1, \ldots, k) \\
& y \in \mathbf{E}.
\end{aligned} \tag{FQP}$$

Theorem 3.4.2. *For any z near $\bar{y}$ and any B positive definite on $G^{\perp}$, the problem (FQP) has a unique minimizer $\hat{y}$ satisfying*

$$\hat{y} - \bar{y} = B^{-1} V (B - H)(z - \bar{y}) + O(|z - \bar{y}|^2)$$

where

$$H = \nabla^2 f(\bar{y}) + \sum_{i=1}^{k} \bar{\lambda}_i \nabla^2 g_i(\bar{y})$$

$$V = I - \nabla G(\nabla G^* B^{-1} \nabla G)^{-1} \nabla G^* B^{-1},$$

with ∇G being the operator defined by

$$\nabla G : \mu \mapsto \sum_{i=1}^{k} \mu_i \nabla g_i(\bar{y}).$$

Proof. See [6, Section 3.2]. □

3.5 Max functions

In this section we carefully analyze how Algorithm 3.2 behaves when applied to a *max function*. Specifically, we assume that the nonsmooth function f has the form

$$f(x) = \max_{i=1,\ldots,k} f_i(x) \tag{3.5.1}$$

for some *structure* functions $f_i : \mathbf{E} \to \mathbf{R}$ for $i = 1, \ldots k$ that are assumed to be $C^{(2)}$-smooth but *unknown*. As previously, we also assume that $r : \mathbf{E} \to \mathbf{R}$ is $C^{(2)}$.

Note that minimizing $F = f + r$ is equivalent to the nonlinear program

$$\begin{aligned} \text{minimize} \quad & t + r(x) \\ \text{subject to} \quad & f_i(x) - t \geq 0 \quad (i = 1, \ldots, k) \\ & x \in \mathbf{E}, \quad t \in \mathbf{R}, \end{aligned} \tag{3.5.2}$$

which could be solved via a standard nonlinear programming algorithm. Our interest, however, in max functions is as local models for more general objective functions, and as a relatively simple test for the general purpose Algorithm 3.2. In this case, we assume that function value, gradient, and Hessian information

of f is returned via an oracle with *no access to the structure functions* f_i. Such a setting destroys any classical approach to solving (3.5.2), since the constraints are only implicitly defined, and cannot be evaluated.

Now, corresponding to the fixed implicit representation

$$F(x) = \max_{i=1,\dots,k} f_i(x) + r(x), \tag{3.5.3}$$

we consider Algorithm 3.2 on a neighbourhood of a *strong nondegenerate* local minimizer $\bar{x}$. Specifically, we assume the strong second-order conditions defined below.

Definition 3.5.1. Given a max function of the form (3.5.3), we say that a point $\bar{x} \in \mathbf{E}$ satisfies the *strong second-order conditions* when the following properties hold.

(i) *Full activity*: the values $f_i(\bar{x})$ are equal for $i = 1, \dots, k$.

(ii) *Independence*: the gradients $\{\nabla f_i(\bar{x}) : i = 1, \dots, k\}$ are affinely independent.

(iii) *Stationarity*: there exists a (necessarily unique) Lagrange multiplier vector $\bar{\lambda} \in \mathbf{R}^k_+$ satisfying $\sum_{i=1}^k \bar{\lambda}_i \nabla f_i(\bar{x}) + \nabla r(\bar{x}) = 0$ and $\sum_{i=1}^k \bar{\lambda}_i = 1$.

(iv) *Second-order sufficiency*: $\sum_{i=1}^k \bar{\lambda}_i \nabla^2 f_i(\bar{x}) + \nabla^2 r(\bar{x})$ is positive definite on the subspace $\left\{z \in \mathbf{E} : \langle \nabla f_i(\bar{x}), z\rangle \text{ equal for } i = 1, \dots, k\right\}$.

(v) *Nondegeneracy*: $\bar{\lambda}_i > 0$ for $i = 1, \dots, k$.

These assumptions closely mirror classical second-order conditions and constraint qualifications in nonlinear programming. Corresponding to the feasible point $(\bar{x}, f(\bar{x}))$ for the problem (3.5.2), full activity amounts to all constraints being active, and independence is the standard linear independence constraint

qualification (LICQ). Stationarity and second-order sufficiency correspond exactly to the analogous nonlinear programming sufficient conditions for $(\bar{x}, f(\bar{x}))$ to be a strict local minimizer. The last condition is frequently known as *strict complementary slackness*, we refer to it as nondegeneracy since (assuming the first three conditions) it is equivalent to $-\nabla r(\bar{x}) \in \mathrm{ri}(\partial f(\bar{x}))$.

We will refer to the disjoint open sets

$$\mathcal{D}_i = \left\{x \in \mathbf{R}^n : f_i(x) > f_j(x) \text{ for all } j \neq i\right\}$$

as *activity regions* of f. Notice that the values, gradients, and Hessians of f and f_i coincide on $\mathcal{D}_i$, and that f is twice continuously differentiable on the open set

$$\mathcal{D} = \bigcup_{i=1}^{k} \mathcal{D}_i.$$

In theory, the algorithm must terminate if a point outside of $\mathcal{D}$ is encountered, since the gradient and Hessian are no longer defined. In practice $\mathbf{E} \setminus \mathcal{D}$ is usually a small set, so this is rarely an issue, especially if computations are performed in inexact or floating-point arithmetic. For the case of a max function under the strong second-order assumptions, each equation $f_i(x) - f_j(x) = 0$ for $j \neq i$ defines a smooth manifold of codimension 1 around $\bar{x}$. Thus $\mathbf{E} \setminus \mathcal{D}$ is contained in a finite union of $(n-1)$-dimensional smooth manifolds, so $\mathcal{D}$ is a dense open set around $\bar{x}$.

Finally, we will only analyze Algorithm 3.2 with the choices of tolerances $\bar{\epsilon} = \bar{\delta} = 0$, so the optimality checks never cause the algorithm to stop.

Full bundles

We consider initializing Algorithm 3.2 with *full* bundle, defined below, contained in $B_\delta(\bar{x})$ for small radius δ. We will prove that the algorithm maintains

bundles satisfying these properties as it proceeds.

Definition 3.5.2. S is a *full bundle* if it can be written in the form

$$S = \{x_1, \ldots, x_k\}$$

where $x_i \in \mathcal{D}_i$ for $i = 1, \ldots, k$.

At the outset of each iteration, we form the linear approximations to f

$$l_s(\,\cdot\,) = f(s) + \langle \nabla f(s), \cdot - s \rangle,$$

and the quadratic approximations to $F = f + r$

$$q_s(\,\cdot\,) = F(s) + \langle \nabla F(s), \cdot - s \rangle + \frac{1}{2} \langle \cdot - s, \nabla^2 F(s)(\cdot - s) \rangle,$$

for $s \in S$. We then compute a multiplier vector by solving

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \Big| \sum_{s \in S} \lambda_s F(s) \Big|^2 \\ \text{subject to} \quad & \sum_{s \in S} \lambda_s = 1 \\ & \lambda \geq 0. \end{aligned} \tag{3.5.4}$$

The next proposition shows that this vector is a good approximation of the optimal Lagrange multiplier $\bar{\lambda}$.

Proposition 3.5.1. *For all small $\delta > 0$ and $\{x_1, \ldots, x_k\} \subset B_\delta(\bar{x})$, the problem*

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \Big| \sum_{i=1}^{k} \lambda_i \left(\nabla f_i(x_i) + \nabla r(x_i) \right) \Big|^2 \\ \textit{subject to} \quad & \sum_{i=1}^{k} \lambda_i = 1 \\ & \lambda \in \mathbf{R}^k \end{aligned} \tag{3.5.5}$$

has a unique minimizer satisfying $\lambda = \bar{\lambda} + O(\delta)$.

Proof. $\lambda \in \mathbf{R}^k$ solves (3.5.5) if and only if $\sum_{i=1}^k \lambda_i = 1$ and there exists some multiplier $\alpha \in \mathbf{R}$ such that

$$\alpha + \Big\langle \nabla f_i(x_i) + \nabla r(x_i), \sum_{j=1}^k \lambda_j (\nabla f_j(x_j) + \nabla r(x_j)) \Big\rangle = 0 \qquad (i = 1, \ldots, k),$$

which is a linear system of the form

$$M(x_1, \ldots, x_k)(\lambda, \alpha) = b$$

for some linear operator M depending smoothly on $x_1, \ldots, x_k$. Clearly, a solution when $x_1 = \ldots = x_k = \bar{x}$ is $(\bar{\lambda}, 0)$. Moreover, we claim that the operator $M(\bar{x}, \ldots, \bar{x})$ is invertible. Indeed, suppose that $M(\bar{x}, \ldots, \bar{x})(\lambda, \alpha) = 0$. Then $\sum_{i=1}^k \lambda_i = 0$ and

$$\begin{aligned} 0 &= \sum_{i=1}^k \lambda_i \Big[\alpha + \Big\langle \nabla f_i(\bar{x}) + \nabla r(\bar{x}), \sum_{j=1}^k \lambda_j (\nabla f_j(\bar{x}) + \nabla r(\bar{x})) \Big\rangle \Big] \\ &= \Big\langle \sum_{i=1}^k \lambda_i \nabla f_i(\bar{x}), \sum_{j=1}^k \lambda_j \nabla f_j(\bar{x}) \Big\rangle. \end{aligned}$$

But then $\lambda = 0$ and hence $\alpha = 0$ by the affine independence of $\{\nabla f_i(\bar{x}) : i = 1, \ldots k\}$, so the result follows from the implicit function theorem. □

Since S is a full bundle, and because of the nondegeneracy assumption, the problems (3.5.4) and (3.5.5) have the same optimal solution for δ sufficiently small.

We now address the quadratic program that computes the new reference point. Since we assume full activity, $\bar{x}$ is a strict local minimizer of the more restrictive problem

$$\begin{aligned} \text{minimize} \quad & t + r(x) \\ \text{subject to} \quad & f_i(x) - t = 0 \quad (i = 1, \ldots, k) \\ & x \in \mathbf{E}, \quad t \in \mathbf{R}. \end{aligned}$$

Therefore given a multiplier $\lambda = \bar{\lambda} + O(\delta)$ computed via (3.5.4), we can apply Theorem 3.4.1 to deduce that the quadratic program

$$\begin{aligned} \text{minimize} \quad & t + \sum_{i=1}^{k} \lambda_i \left[\langle \nabla r(x_i), x - x_i \rangle + \frac{1}{2} \langle x - x_i, (\nabla^2 f(x_i) + \nabla^2 r(x_i))(x - x_i) \rangle \right] \\ \text{subject to} \quad & f(x_i) + \langle \nabla f(x_i), x - x_i \rangle - t = 0 \quad (i = 1, \ldots, k) \\ & x \in \mathbf{E}, \quad t \in \mathbf{R}, \end{aligned}$$

has a unique minimizer $\hat{x} = \bar{x} + O(\delta^2)$. This $\hat{x}$ is exactly the new reference point computed by the algorithm, since for full bundles S, the problem is equivalent to our quadratic subproblem

$$\begin{aligned} \text{minimize} \quad & \sum_{s \in S} \lambda_s q_s(x) \\ \text{subject to} \quad & l_s(x) \text{ equal for } s \in S \\ & x \in \mathbf{E}. \end{aligned}$$

In particular the subproblem is never unbounded below due to our second-order assumptions.

Finally, we address the reference set update procedure, replaces a single point $s \in S$ with the new reference point $\hat{x}$ so as to minimize the optimality measure

$$\Theta(S) = \min \left| \text{conv}(\nabla F(S)) \right|.$$

This update procedure results in another full bundle, which we show by first proving a simple tool.

Proposition 3.5.2. *There exist constants $\epsilon, \delta > 0$ such that for all $S \subset B_\delta(\bar{x})$, all indices $i = 1, \ldots, k$ and all points $\hat{x} \in \mathcal{D}_i \cap B_\delta(\bar{x})$,*

$$\min \left| \text{conv} \left(\{ \nabla F(x_j) : j \neq p \} \cup \nabla F(\hat{x}) \right) \right| > \epsilon$$

for all $p \neq i$.

Proof. Suppose the result does not hold. Then there exists some $p \neq i$ and sequences $x_j^r \in \mathcal{D}_j$ for $j = 1, \ldots, k$ and $\hat{x}^r \in \mathcal{D}_i$ all approaching $\bar{x}$ such that

$$\left| \mu^r (\nabla f(\hat{x}^r) + \nabla r(\hat{x}^r)) + \sum_{j \neq p} \lambda_j^r (\nabla f(x_j) + \nabla r(x_j)) \right| \to 0$$

as $r \to \infty$. By continuity of the gradients and dropping to a convergence subsequence it follows that

$$0 \in \operatorname{conv}\left(\{ \nabla f_j(\bar{x}) + \nabla r(\bar{x}) : j \neq p \} \right),$$

which contradicts our nondegeneracy assumption. □

With this we can deduce that the algorithm maintains full bundles as it progresses.

Corollary 3.5.1. *For all small $\delta > 0$, full bundles $S \subset B_\delta(\bar{x})$, and new reference point $\hat{x} \in \mathcal{D} \cap B_\delta(\bar{x})$, there is a unique reference point $s \in S$ that minimizes*

$$\min \left| \operatorname{conv}\left(\{ \nabla F(s') : s \neq s' \in S \} \cup \nabla F(\hat{x}) \right) \right|$$

and the set $(S \setminus s) \cup \{\hat{x}\}$ is a full bundle.

Proof. Without loss of generality suppose $\hat{x} \in \mathcal{D}_i$. Let δ be sufficiently small so that Proposition 3.5.2 holds. Then by continuity, and shrinking δ if necessary, Proposition 3.5.1 implies that

$$\min \left| \operatorname{conv}\left(\{ \nabla F(x_j) : x_j \in S \} \right) \right| < \epsilon.$$

Again shrinking δ if necessary, we can assume $|\nabla f(x_i) - \nabla f(\hat{x})| < \epsilon$, so

$$\min \left| \operatorname{conv}\left(\{ \nabla F(x_j) : j \neq i \} \cup \nabla F(\hat{x}) \right) \right| < \epsilon.$$

Then Proposition 3.5.2 implies that $s = x_i$ is the unique minimizer. □

To summarize, we have proved the following.

Theorem 3.5.1. *Given a max function of the form*

$$F(x) = \max_{i=1,\dots,k} f_i(x) + r(x)$$

with $\bar{x}$ satisfying the strong second-order conditions, there exists a constants $\delta, M > 0$ such that starting from any full bundle $S \subset B_\delta(\bar{x})$, Algorithm 3.2 generates a new point $\hat{x}$ satisfying

$$|\hat{x} - \bar{x}| \leq M \max_{s \in S} |s - \bar{x}|^2.$$

Assuming $\hat{x} \in \mathcal{D}$, the algorithm replaces with $\hat{x}$ the reference point in S from the same activity region, generating a new full bundle.

While this guarantees the algorithm cannot diverge, we cannot yet guarantee convergence to $\bar{x}$, since it is conceivable the algorithm never updates the reference point $\arg\max\{|s - \bar{x}| : s \in S\}$. To guarantee that the sequence of bundles shrink to $\bar{x}$, we must impose a strong convexity assumption on the nonsmooth function f. We first develop a simple tool for sequences of positive numbers.

Lemma 3.5.1. *Let $\alpha, M > 0$. Consider any sequence of vectors $z \in \mathbf{R}^k_+$ such that each successive pair z, z' in the sequence has the property that there exists i such that*

$$z_i \geq \alpha \|z\|_{\max} \quad \textit{and} \quad z'_i \leq M \|z\|^2_{\max},$$

and that $z'_j = z_j$ for $j \neq i$. Then providing that the initial vector z satisfies $\|z\|_{\max} < \frac{1}{M}$, the sequence converges to zero at a k-step quadratic rate.

Proof. Given the initial bound, $\|z\|_{\max}$ is clearly nondecreasing. Now, fix an arbitrary vector in the sequence z_{old}, and fix $\gamma = \|z_{\text{old}}\|_{\max}$. At the next and every

subsequent iteration, some element z_i is set to a value in the interval $[0, M\gamma^2]$. Therefore the element z_i cannot be updated again unless we have that

$$\|z\|_{\max} \leq \frac{M}{\alpha}\gamma^2,$$

since otherwise

$$\alpha \|z\|_{\max} > M\gamma^2 \geq z_i,$$

which violates the rules of the sequence. Therefore after at most n iterations, we arrive at a vector z_{new} such that

$$\|z_{\text{new}}\|_{\max} \leq \frac{M}{\alpha} \|z_{\text{old}}\|_{\max}^2.$$

□

Theorem 3.5.2. *Given a max function representation of the objective*

$$F(x) = \max_{i=1,\ldots,k} f_i(x) + r(x)$$

for $C^{(2)}$-smooth r, f_i, suppose the point $\bar{x}$ satisfies the strong second-order conditions of Definition 3.5.1, and that the Hessians $\nabla^2 f_i(\bar{x})$ are positive definite. Then there exists $\delta > 0$ such that starting from a full bundle $S \subset B_\delta(\bar{x})$, as long as a point outside of $\mathcal{D}$ is not encountered, Algorithm 3.2 with tolerances $\bar{\epsilon} = \bar{\delta} = 0$ generates a sequence of full bundles that converge k-step quadratically to $\bar{x}$.

Proof. There exists $\rho > 0$ such that each function f_i is ρ-strongly convex on $B_\delta(\bar{x})$. Shrinking δ if necessary, we can apply Theorem 3.5.1 for some full bundle $S \subset B_\delta(\bar{x})$. Then we have that every iteration of the algorithm replaces a reference point x_i with a new point $\hat{x} \in \mathcal{D}_i \cap B_\delta(\bar{x})$ so that

$$|\hat{x} - \bar{x}| \leq M \max_{j=1,\ldots,k} |x_j - \bar{x}|^2.$$

Since f_i is $C^{(2)}$, there exists $L \geq \rho$ such that

$$f_i(x) \leq l_i(x) + \frac{L}{2}|x - x_i|^2 \quad \text{for all } x \in B_\delta(\bar{x}).$$

On the other hand, by strong convexity

$$f_j(x) \geq l_j(x) + \frac{\rho}{2}|x - x_j|^2 \quad \text{for all } x \in B_\delta(\bar{x}) \qquad (j = 1 \ldots, k).$$

Also, by construction $f_i(\hat{x}) \geq f_j(\hat{x})$ and $l_i(\hat{x}) = l_j(\hat{x})$ for all j. Altogether we have that for all $j = 1, \ldots, k$,

$$l_i(\hat{x}) + \frac{L}{2}|\hat{x} - x_i|^2 \;\geq\; f_i(\hat{x}) \;\geq\; f_j(\hat{x}) \;\geq\; l_j(\hat{x}) + \frac{\rho}{2}|\hat{x} - x_j|^2,$$

which yields

$$|\hat{x} - x_i| \geq \kappa\,|\hat{x} - x_j|$$

for $\kappa = \sqrt{\frac{\rho}{L}} \leq 1$. Letting $\beta = \max_{j=1,\ldots,k}|x_j - \bar{x}|^2$, this implies that

$$\begin{aligned}
\kappa\,|x_j - \bar{x}| &\leq \kappa\,|x_j - \hat{x}| + \kappa\,|\hat{x} - \bar{x}| \\
&\leq |\hat{x} - x_i| + \kappa M\beta^2 \\
&\leq |\hat{x} - \bar{x}| + |x_i - \bar{x}| + \kappa M\beta^2 \\
&\leq |x_i - \bar{x}| + (1 + \kappa)M\beta^2.
\end{aligned}$$

Maximizing over j yields

$$|x_i - \bar{x}| \geq \kappa\beta - (1 + \kappa)M\beta^2.$$

For any $c > 1$, letting $\beta \leq \left(1 - \frac{1}{c}\right)\frac{\kappa}{1+\kappa}\frac{1}{M} < \frac{1}{M}$ (by shrinking δ if necessary) implies that $|x_i - \bar{x}| \geq \frac{\kappa}{c}\beta$. To summarize, the new reference point $\hat{x}$ and old reference point x_i satisfy the two inequalities

$$|x_i - \bar{x}| \geq \frac{\kappa}{c}\max_j |x_j - \bar{x}| \quad \text{and} \quad |\hat{x} - \bar{x}| \leq M \max_j |x_j - \bar{x}|^2.$$

Applying Lemma 3.5.1 with

$$z_j = |x_j - \bar{x}| \qquad (j = 1, \ldots, k)$$

completes the proof. □

The assumption that the Hessians $\nabla^2 f_i(\bar{x})$ are positive definite may seem stringent, but the smooth-nonsmooth model (3.5.3) is quite flexible in practice. In the following section we will show how to apply a bundle Newton algorithm to *weakly convex* functions.

3.6 Weakly convex minimization

Recall a function $F : \mathbf{E} \to \mathbf{R}$ is weakly convex if $F + \frac{\eta}{2}|\cdot|^2$ is convex for η sufficiently large (we can assume strong convexity by increasing η if necessary). Then, assuming that F is twice continuously differentiable on the set $\mathcal{D}$, we can apply Algorithm 3.2 to the problem of minimizing F by defining

$$f = F + \frac{\eta}{2}|\cdot|^2, \qquad r = -\frac{\eta}{2}|\cdot|^2.$$

With some trivial simplifications we arrive at the following algorithm and convergence result.

Corollary 3.6.1. *Given a max function representation of the objective*

$$F(x) = \max_{i=1,\ldots,k} f_i(x)$$

for $C^{(2)}$-smooth f_i, suppose the point $\bar{x}$ satisfies the strong second-order conditions of Definition 3.5.1. Then there exists $\delta > 0$ such that starting from a full bundle $S \subset B_\delta(\bar{x})$, as long as a point outside of $\mathcal{D}$ is not encountered, Algorithm 3.3 with tolerances $\bar{\epsilon} = \bar{\delta} = 0$ and sufficiently large weak convexity parameter $\eta > 0$ generates a sequence of full bundles that converge k-step quadratically to $\bar{x}$.

Algorithm 3.3: Local Newton algorithm to minimize weakly convex F

Input : Bundle $S \subset \mathcal{D}$, tolerances $\bar{\epsilon}, \bar{\delta} \geq 0$, weak convexity parameter $\eta \geq 0$;
while diam $S > \bar{\delta}$ and $\Theta(S) > \bar{\epsilon}$ **do**
  for $s \in S$ **do**
    $l_s(\cdot) = F(s) + \langle \nabla F(s), \cdot - s \rangle + \eta \langle s, \cdot \rangle - \frac{\eta}{2} |s|^2$;
    $q_s(\cdot) = F(s) + \langle \nabla F(s), \cdot - s \rangle + \frac{1}{2} \langle \cdot - s, \nabla^2 F(s)(\cdot - s) \rangle$;
  end
  Choose $\lambda \in \Delta_S$ to minimize $\left| \sum_{s \in S} \lambda_s \nabla F(s) \right|$;
  if $\min\left\{ \sum_{s \in S} \lambda_s q_s(x) : l_s(x) \textit{ equal for all } s \in S \right\} = -\infty$ **then**
    Stop;
  else
    Choose optimal $\hat{x}$;
    if $x \notin \mathcal{D}$ **then**
      Stop;
    else
      Choose $s \in S$ minimizing $\Theta\big((S \setminus s) \cup \hat{x}\big)$;
      $S \leftarrow (S \setminus s) \cup \hat{x}$;
    end
  end
end

Proof. Define the functions

$$f_i = g_i + \frac{\eta}{2} |x|^2 \quad (i = 1, \ldots, k) \quad \text{and} \quad r = -\frac{\eta}{2} |x|^2.$$

Choosing

$$\eta > \max_{i=1,\ldots,k} \lambda_{\min}(\nabla^2 F_i(\bar{x}))$$

ensures that each function f_i is strongly convex around $\bar{x}$ with $\nabla^2 f_i(\bar{x})$ positive definite. Also, it is easy to check that $\bar{x}$ satisfies the strong second-order conditions for $\max_{i=1,\ldots,k}\{f_i(x) - r(x)\}$, and that the activity regions for f and F coincide. Thus we can apply Theorem 3.5.2 to yield the result. □

We remark that this approach seems to share some similarities with *trust-region* Newton methods, which can also be interpreted as shifting the eigenval-

ues of the Hessian to enforce convexity in the model (see for example [106, Theorem 4.1]). It would be interesting to explore this connection in future work.

3.7 Numerical experiments

This section illustrates the local bundle Newton algorithm on several nonsmooth objective functions. These simple experiments are meant as a proof of concept, not a comprehensive numerical study. Nonetheless, the results appear clearly promising enough to invite future research.

3.7.1 Practical considerations

None of the stopping criteria were implemented, and the algorithm was terminated manually when rounding error prevented any further progress.

Choosing an initial bundle

In each experiment, a standard global nonsmooth optimization method was used to generate a finite set of points $\Omega \subset \mathcal{D}$ near a minimizer $\bar{x}$, and used the corresponding gradients (normalized to unit length) to estimate the dimension of the subdifferential $\partial f(\bar{x})$ and hence choose the bundle size k, as discussed in Section 3.2.5. A heuristic subset selection procedure [56, Algorithm 5.5.1], described below, was then used to choose a set of k points in Ω with robustly affinely independent gradients to form the initial bundle.

For convex problems, we use the simple "Bundle Method with Multiple Cuts" [44], described in Algorithm 3.5. In the nonconvex case, we use BFGS [82].

Algorithm 3.4: Bundle initialization heuristic

Input : Set of points $\Omega = \{x_1, \ldots, x_m\} \subset \mathbf{R}^n$, singular value threshold $\bar{\sigma}$;

1. Form the matrix $G = [g_1 \cdots g_m]$ with columns

$$g_i = \begin{pmatrix} |\nabla f(x_i)|^{-1} \nabla f(x_i) \\ 1 \end{pmatrix};$$

2. Compute SVD and determine numerical rank of G,

$$U^T G V = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n),$$
$$r = \max\{1 \leq i \leq m : \sigma_i > \bar{\sigma}\};$$

3. Apply QR with column pivoting to the first r columns of V, $V_r = [v_1 \cdots v_r]$,

$$Q^{\mathsf{T}} V_r P = R;$$

4. Let p be the list that results from permuting the list $\{1, \ldots, n+1\}$ according to the permutation matrix P. Return the bundle of points

$$\{x_j : j \in \{p_1, \ldots, p_r\}\};$$

For the bundle method, Ω is chosen to be the set of points whose cutting planes were strongly active in the final iteration. That is,

$$\Omega \; = \; \{s \in S : \alpha_s > 0\},$$

where α_s is the dual variable associated with cutting plane l_s. For BFGS, Ω is the final $2n$ iterates.

Solving the quadratic subproblems

The algorithm involves two quadratic programming subproblems. The first involves computing the optimality measure $\Theta(S)$, which amounts to choosing the smallest element in the convex hull of vectors $\{\nabla f(s) : s \in S\}$. This was

Algorithm 3.5: Multiple cut bundle method to minimize convex f

Input : Initial bundle $S \subset \mathbf{R}^n$, initial center $z \in S$, stopping tolerance $\bar{\epsilon}$, proximal parameter $\rho > 0$, sufficient decrease parameter $\beta \in (0, 1)$;

for iteration $= 1, 2, 3, \ldots$ **do**

  for $s \in S$ **do**

    $g_s \in \partial f(s)$;

    $l_s(\cdot) = f(s) + \langle g_s, \cdot - s \rangle$;

  end

  Choose $\hat{x}$ minimizing $\max_{s \in S} l_s(\cdot) + \frac{\rho}{2} |\cdot - z|^2$;

  if $f(z) - \max_{s \in S} l_s(\hat{x}) \leq \bar{\epsilon}$ **then**

    Stop: nearly optimal;

  else

    if $f(\hat{x}) \leq f(z) - \beta(f(z) - \max_{s \in S} l_s(\hat{x}))$ **then**

      $z \leftarrow \hat{x}$ (serious step);

    else

      $z \leftarrow z$ (null step);

    end

  end

  $S \leftarrow S \cup \{\hat{x}\}$;

end

implemented as the quadratic program

$$\begin{aligned} \underset{\lambda \in \mathbf{R}^{|S|}}{\text{minimize}} \quad & \frac{1}{2} \Big| \sum_{s \in S} \lambda_s \nabla f(s) \Big|^2 \\ \text{subject to} \quad & \sum_{s \in S} \lambda_s = 1 \\ & \lambda \geq 0, \end{aligned}$$

and solved in Gurobi. For the equality-constrained quadratic programs,

$$\begin{aligned} \underset{t \in \mathbf{R},\ x \in \mathbf{E}}{\text{minimize}} \quad & \sum_s \lambda_s q_s(x) \\ \text{subject to} \quad & l_s(x) \text{ equal for all } s \in S, \end{aligned}$$

we simply solve the linear optimality conditions directly:

$$\begin{aligned} \sum_{s\in S} \lambda_s \nabla^2 f(s)(x-s) + \sum_{s\in S} \mu_s \nabla f(s) &= 0 \\ \sum_{s\in S} \mu_s &= 1 \\ l_s(x) - t &= 0 \quad (s \in S). \end{aligned} \tag{3.7.1}$$

The x variable of the solution is then our Newton iterate $\hat{x}$.

3.7.2 Illustrative examples

Strongly convex problems

Our first experiment is to minimize max functions of the form

$$f(x) = \max_{i=1,\dots k} \left\{ g_i^T x + \frac{1}{2} x^T H_i x + \frac{c_i}{24} |x|^4 \right\} \qquad (x \in \mathbf{R}^n) \tag{3.7.2}$$

for $1 \le k \le n+1$. Consider randomly generated positive constants c_i, symmetric positive definite matrices H_i, and affinely independent vectors g_i satisfying $\sum_i \lambda_i g_i = 0$ for some λ randomly sampled in $\{\lambda > 0 : \sum_i \lambda_i = 1\}$. Then

$$0 \in \partial f(0) = \operatorname{conv}\{g_i : 1 \le i \le k\}$$

so f is nonsmooth at the minimizer of 0, and it is clear that 0 satisfies the strong second-order conditions. This structure of f is unknown to the algorithms. Access is limited to a black box procedure that returns the active function value, gradient, and Hessian.

In random trials for dimension $n = 50$, the bundle method Algorithm 3.5 was applied in a first phase, with parameters $\rho = 1$ and $\beta = 10^{-5}$, and starting point $z = (1, \dots, 1)$. The stopping tolerance was $\bar{\epsilon}$ set to 10^{-6}, at which point we initialize and switch to Algorithm 3.1. Results for a number of random trials are shown

in Figures 3.1 and 3.2. For the bundle method, we observe a roughly linear rate of convergence of function values to zero that is proportional to the "degree of nonsmoothness" k. Switching to the bundle Newton algorithm results in rapid convergence of function values to zero.

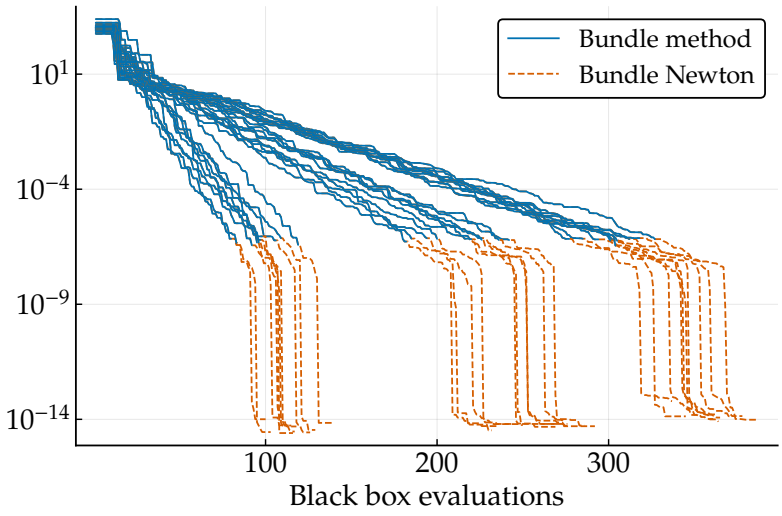
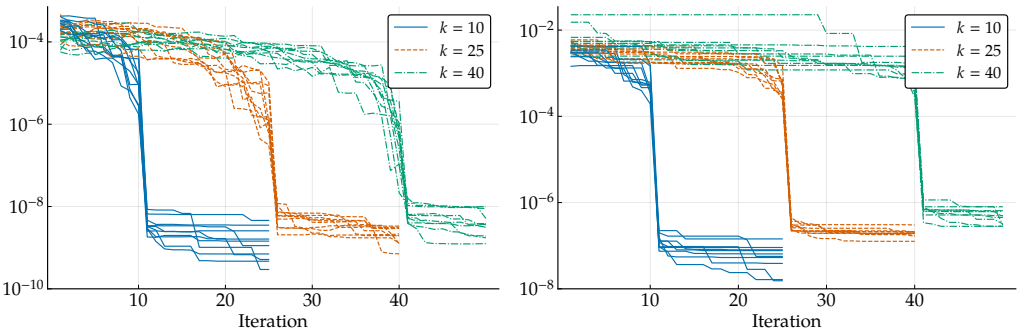


Figure 3.1: Best function value found for the bundle method and bundle Newton algorithm against number of black box evaluations for random max functions (3.7.2) for $k = 10, 25, 40$ in dimension $n = 50$.



(a) $\Theta(S)$ (b) diam S

Figure 3.2: Optimality measures against iteration count for the bundle Newton algorithm for random max functions (3.7.2) in dimension $n = 50$.

Nonconvex problems

As a nonconvex test, consider functions of the form

$$f(x) = \sum_{i=1}^{k} \left| g_i^T x + \frac{1}{2} x^T H_i x + \frac{c_i}{24} |x|^4 \right| \qquad (x \in \mathbf{R}^n) \tag{3.7.3}$$

for $1 \leq k \leq n+1$, with constants c_i, vectors g_i and matrices H_i randomly generated as in the previous experiment. As usual, access to f is limited to a black box that returns function values, gradients, and Hessians.

In random trials for dimension $n = 50$, nonsmooth BFGS was applied in a first phase until a breakdown occurred due to numerical instability (as usual with this method [82]). At this point we switch to the Algorithm 3.3 with weak convexity parameter dynamically chosen as

$$\eta = \max_{s \in S} \lambda_{\max} \left(-\nabla^2 f(s) \right)$$

at each iteration. We observe that the algorithm is quickly able to significantly improve the solution returned by BFGS.

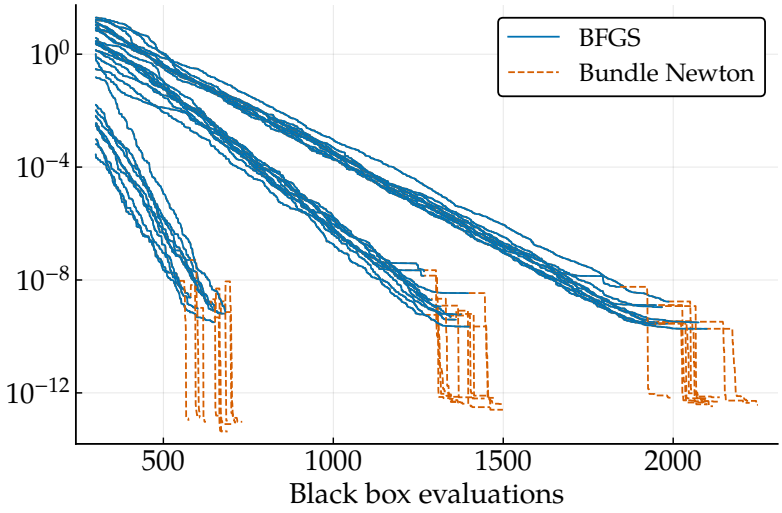
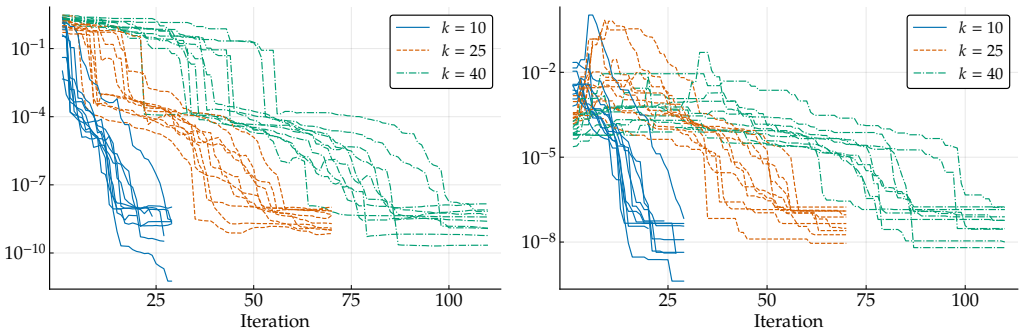


Figure 3.3: Best function value found for BFGS and the bundle Newton algorithm against number of black box evaluations on random Euclidean sum functions (3.7.3) for $k = 10, 25, 40$ in dimension $n = 50$



(a) $\Theta(S)$ (b) diam S

Figure 3.4: Optimality measures against iteration count for the bundle Newton algorithm for random Euclidean sum functions (3.7.3) in dimension $n = 50$.

3.8 First-order analogues

The Newton philosophy that we explored in this chapter is suggestive even in the more usual setting where Hessians are unavailable. One straightforward first-order analogue of our second-order methods replaces the Hessians by a fixed multiple of the identity matrix, resulting in quadratic subproblem objectives of the form

$$\sum_{s \in S} \lambda_s \left[f(s) + \nabla f(s)^T (\cdot - s) + \frac{\rho}{2} |\cdot - s|^2 \right]$$

for some suitable choice of $\rho > 0$.

This simple implementation is effective on max functions: a simple illustration is the nonsmooth Rosenbrock function

$$f(x) = \frac{1}{8}(1 - x_1)^2 + |x_2 - x_1^2|. \tag{3.8.1}$$

It is easy to verify that $f + |\cdot|^2$ is strongly convex. Moreover, relative to the set $\{x \in \mathbf{R}^2 : x_1^2 = x_2\}$, f behaves as the smooth univariate function $\frac{1}{8}(1 - \cdot)^2$, suggesting parameter choices of $\eta = 2$ and $\rho = \frac{1}{4}$ in a first-order implementation. Local convergence (after using BFGS to initialize a bundle of size $k = 2$) is shown in

Algorithm 3.6: First-order algorithm to minimize weakly convex f

Input : Bundle $S \subset \mathcal{D}$, tolerances $\bar{\epsilon}, \bar{\delta} \geq 0$, weak convexity parameter $\eta \geq 0$, scaling parameter $\rho > 0$;
while diam $S > \bar{\delta}$ and $\Theta(S) > \bar{\epsilon}$ **do**
  for $s \in S$ **do**
    $l_s(\cdot) = f(s) + \langle \nabla f(s), \cdot - s \rangle + \eta \langle s, \cdot \rangle - \frac{\eta}{2} |s|^2$;
    $q_s(\cdot) = f(s) + \langle \nabla f(s), \cdot - s \rangle + \frac{\rho}{2} |\cdot - s|^2$;
  end
  Choose $\lambda \in \Delta_S$ to minimize $\left| \sum_{s \in S} \lambda_s \nabla f(s) \right|$;
  Choose $\hat{x} \in \arg\min \left\{ \sum_{s \in S} \lambda_s q_s(x) : l_s(x) \text{ equal for all } s \in S \right\}$;
  if $x \notin \mathcal{D}$ **then**
    Stop;
  else
    Choose $s \in S$ minimizing $\Theta\big((S \setminus s) \cup \hat{x}\big)$;
    $S \leftarrow (S \setminus s) \cup \hat{x}$;
  end
end

Figure 3.5, where we observe an overall linear rate. In the case of a max function, we have the following extension of Theorem 3.5.1.

Theorem 3.8.1. *Given a max function of the form*

$$f(x) = \max_{i=1,\ldots,k} f_i(x)$$

with $\bar{x}$ satisfying the strong second-order conditions, there exists a constant $\delta > 0$ such that starting from any full bundle $S \subset B_\delta(\bar{x})$, Algorithm 3.6 with $\eta = 0$ generates a new point $\hat{x}$ satisfying

$$\hat{x} - \bar{x} = \mathrm{Proj}_L \left(\frac{1}{\rho} \sum_{s \in S} \bar{\lambda}_s \big(\rho I - \nabla^2 f_s(\bar{x})\big)(s - \bar{x}) \right) + O\left(\max_{s \in S} |s - \bar{x}|^2 \right),$$

where L is the subspace $\{z \in \mathbf{R}^n : \langle \nabla f_i(\bar{x}), z \rangle$ equal for all $i\}$. Assuming $\hat{x} \in \mathcal{D}$, the algorithm replaces with $\hat{x}$ the reference point in S from the same activity region, generating a new full bundle.

Proof. Fix the full bundle $S = \{y_1, \ldots, y_k\}$. We apply Theorem 3.4.2 to the con-

sensus form of our quadratic subproblem

$$\begin{aligned} \text{minimize} \quad & t + \frac{\rho}{2} \sum_{i=1}^{k} \lambda_i \, |x_i - y_i|^2 \\ \text{subject to} \quad & f_i(y_i) + \langle \nabla f_i(y_i), x_i - y_i \rangle - t = 0 \quad (i = 1, \ldots, k) \\ & x_i - z = 0 \quad (i = 1, \ldots, k) \\ & x_1, \ldots, x_k, z \in \mathbf{E}. \end{aligned}$$

with the operator B defined as $(x_1, \ldots, x_k, z) \mapsto \rho(\lambda_1 x_1, \ldots, \lambda_k x_k, 0)$. Then noticing that $B^{-1}V$ is simply composed of multiple projections onto L, we deduce that

$$z - \bar{x} = \text{Proj}_L \left(\frac{1}{\rho} \sum_{i=1}^{k} \bar{\lambda}_i \big(\rho(\lambda_i / \bar{\lambda}_i) I - \nabla^2 f_i(\bar{x}) \big)(y_i - \bar{x}) \right) + O\Big(\max_{s \in S} |y_i - \bar{x}|^2 \Big).$$

Applying Proposition 3.5.1 and Corollary 3.5.1 completes the result. □

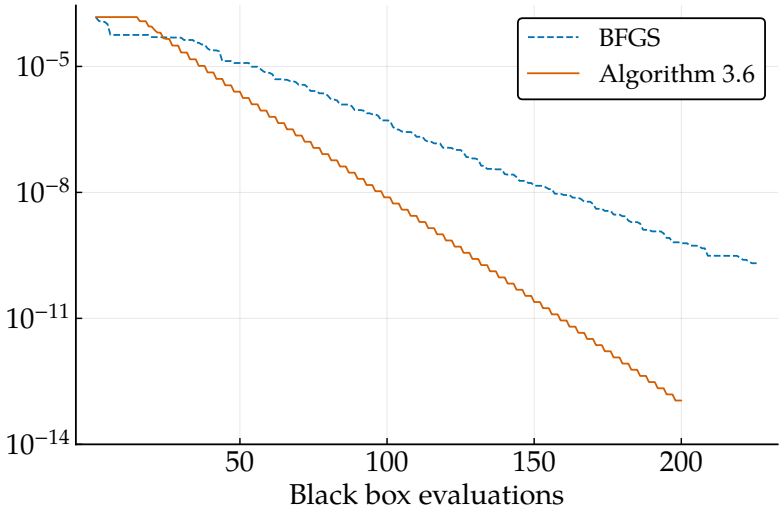


Figure 3.5: Best function value found for BFGS and Algorithm 3.6 against number of black box evaluations on the nonsmooth Rosenbrock function (3.8.1).

To complete a local convergence proof, and exactly characterize the linear convergence rate, an appropriate analogue of Lemma 3.5.1 should be developed. How to do so is not immediately obvious, and is left as a topic for future research.

3.9 Globalization

Since our method depends strongly on the local geometry of the objective function around a minimizer, the results in this chapter have been purely local, with the algorithms crucially depending on an initial full bundle being available. One potentially promising route to a more robust global algorithm is to envision a two-stage approach in the vein of *sequential linear-quadratic programming* [106]. In SLQP, a linear program is solved to estimate the set of active constraints, which are then fixed in an equality-constrained quadratic subproblem. We have experimentally observed that a similar strategy can be effective for convex max functions, with a second linear subproblem introduced to control the size and elements of the bundle, alongside a simple backtracking line search to choose the scaling parameter in a first-order implementation. A more comprehensive investigation is an ongoing study.

An early version of this work considered the algorithmic scheme described below for minimizing max functions assuming each $f_i : \mathbf{E} \to \mathbf{R}$ is convex and L-Lipschitz continuous. (The extension to smooth-nonsmooth sums (3.3.1) is straightforward.)

This algorithm can be analyzed in the framework of the model-based minimization schemes of Davis and Drusvyatskiy [36], of which the following analysis is essentially derived from. Notice that when a full bundle S is sufficiently close to a strong minimizer $\bar{x}$, we have for all $t \geq 1$ that

$$g_t(x) \leq f(x) \quad \text{for all } x,$$
$$g_t(x_t) = f(x_t).$$

The first inequality follows from convexity, and the second from the fact that the

Algorithm 3.7: First-order model-based algorithm to minimize f

Input : Full bundle $S \subset \mathcal{D}$, $x_0 \in S$, step size sequence $(\rho_t) > 0$;
for $t = 0, 1, \ldots$ **do**
  for $s \in S$ **do**
    $l_s(\cdot) = f(s) + \langle \nabla f(s), \cdot - s \rangle$;
  end
  $g_t(\cdot) = \max_{s \in S} l_s(\cdot)$;
  $x_{t+1} = \arg\min_x g_t(x) + \frac{\rho_t}{2} |x - x_t|^2$;
  if $x_{t+1} \notin \mathcal{D}$ **then**
    Stop;
  else
    Choose $s \in S$ minimizing $\Theta\big((S \setminus s) \cup x_{t+1}\big)$;
    $S \leftarrow (S \setminus s) \cup x_{t+1}$;
  end
end

bundle update rule maintains full bundles (Corollary 3.5.1).

Fixing arbitrary $\gamma > 0$, a convergence rate of Algorithm 3.7 can be given in terms of the *Moreau envelope*

$$f^\gamma(x) = \inf_{y \in \mathbf{E}} \Big\{ f(y) + \frac{\gamma}{2} |y - x|^2 \Big\},$$

a smoothed version of f. It is well known from convex analysis that the Moreau envelope is continuously differentiable, with gradient in terms of the proximal point mapping given by

$$\nabla f^\gamma(x) = \gamma \big(x - \operatorname{prox}_{\gamma^{-1} f}(x) \big).$$

When this gradient is small, x is close to strong minimizer $\bar{x}$ [43].

Defining the exact proximal point

$$\hat{x}_t = \operatorname{prox}_{\gamma^{-1} f}(x_t) = \arg\min_{x \in \mathbf{E}} \Big\{ f(x) + \frac{\gamma}{2} |x - x_t|^2 \Big\},$$

by strong convexity of the subproblems and the properties of g_k we have that

$$\begin{aligned}
&\frac{\rho_t}{2}\left|x_{t+1}-x_t\right|^2+\frac{\rho_t}{2}\left|\hat{x}_t-x_{t+1}\right|^2-\frac{\rho_t}{2}\left|\hat{x}_t-x_t\right|^2 \\
&\leq g_t(\hat{x}_t)-g_t(x_{t+1}) \\
&\leq f(\hat{x}_t)-g_t(x_{t+1}) \\
&\leq f(x_{t+1})+\frac{\gamma}{2}\left|x_{t+1}-x_t\right|^2-\frac{\gamma}{2}\left|\hat{x}_t-x_k\right|^2-g_t(x_{t+1}).
\end{aligned}$$

Rearranging and using that f and hence also g_t are L-Lipschitz yields

$$\begin{aligned}
\left|x_{t+1}-\hat{x}_t\right|^2 &\leq \frac{\rho_t-\gamma}{\rho_t}\left|\hat{x}_t-x_t\right|^2+\frac{\gamma-\rho_t}{\rho_t}\left|x_{t+1}-x_t\right|^2+\frac{2}{\rho_t}\big(f(x_{t+1})-g_t(x_{t+1})\big) \\
&\leq \frac{\rho_t-\gamma}{\rho_t}\left|\hat{x}_t-x_t\right|^2+\frac{\gamma-\rho_t}{\rho_t}\left|x_{t+1}-x_t\right|^2+\frac{2}{\rho_t}2L\left|x_{t+1}-x_t\right|.
\end{aligned}$$

Supposing without loss of generality that $\rho_k > \gamma$, by maximizing the above over $|x_{t+1} - x_t|$ we deduce that

$$\left|x_{t+1}-\hat{x}_t\right|^2 \leq \frac{\rho_t-\gamma}{\rho_t}\left|\hat{x}_t-x_t\right|^2+\frac{4L^2}{\rho_k(\rho_k-\gamma)}.$$

Therefore by definition,

$$\begin{aligned}
f^\gamma(x_{t+1}) &\leq f(\hat{x}_t)+\frac{\gamma}{2}\left|\hat{x}_t-x_{t+1}\right|^2 \\
&\leq f(\hat{x}_t)+\frac{\gamma}{2}\left(\frac{\rho_t-\gamma}{\rho_t}\left|\hat{x}_t-x_t\right|^2+\frac{4L^2}{\rho_k(\rho_k-\gamma)}\right) \\
&= f^\gamma(x_t)-\frac{\gamma^2}{2\rho_t}\left|\hat{x}_t-x_t\right|^2+\frac{2\gamma L^2}{\rho_k(\rho_k-\gamma)}.
\end{aligned}$$

Fixing arbitrary $T > 0$ and iterating yields

$$\min f \;\leq\; f^\gamma(x_{T+1}) \;\leq\; f^\gamma(x_0)-\frac{\gamma}{2}\sum_{t=0}^{T}\frac{\gamma}{\rho_t}\left|\hat{x}_t-x_t\right|^2+2\gamma L^2\sum_{t=0}^{T}\frac{1}{\rho_k(\rho_k-\gamma)},$$

and by rearranging we have that

$$\sum_{t=0}^{T}\frac{\gamma}{\rho_t}\left|\hat{x}_t-x_t\right|^2 \leq \frac{2(f^\gamma(x_0)-\min f)}{\gamma}+4L^2\sum_{k=0}^{T}\frac{1}{\rho_k(\rho_k-\gamma)}.$$

Recognizing that

$$|\hat{x}_t - x_t|^2 = \frac{1}{\gamma^2} |\nabla f^\gamma(x_t)|^2,$$

we arrive at the inequality

$$\min_{t=0,\dots,T} |\nabla f^\gamma(x_t)|^2 \leq \frac{2\gamma(f^\gamma(x_0) - \min f) + 4\gamma^2 L^2 + \sum_{t=0}^T \frac{1}{\rho_t(\rho_t - \gamma)}}{\sum_{t=0}^T \frac{\gamma}{\rho_t}}.$$

In particular, choosing $\rho_t = \gamma + \alpha\sqrt{t+1}$ for any $\alpha > 0$ yields the complexity guarantee

$$\min_{t=0,\dots,T} |\nabla f^\gamma(x_t)|^2 \leq \left(2(f^\gamma(x_0) - \min f) + 4\gamma L^2 \alpha^2\right) \left(\frac{\gamma}{T+1} + \frac{1}{\alpha\sqrt{T+1}}\right). \quad (3.9.1)$$

Since we essentially do not use the special finite max structure of f, this rate is not very useful practically, being no better than a simple subgradient method. It would be interesting to investigate if this proof technique based on Moreau envelopes could be specialized to derive new convergence rates for more complex bundle algorithms on functions with finite max structure.

CHAPTER 4

ACTIVE-SET NEWTON METHODS

4.1 Introduction

For the one-dimensional equation $p(x) = 0$, Newton's method has a simple geometric motivation. At an approximate solution x, we intersect the tangent line to the graph at $(x, p(x))$ with the x-axis, and take the intersection point as an improved approximate solution. One key ingredient for rapid convergence to solution $\bar{x}$ is that the derivative $p'(\bar{x})$ is well-defined and nonzero, in which case the Newton iteration

$$x \leftarrow x - \frac{p(x)}{p'(x)}$$

converges quadratically to $\bar{x}$. If this doesn't hold, Newton's method may converge more slowly, such as linearly on the function $p(x) = x^2$, or even diverge, such as when applied to $p(x) = \sqrt[3]{x}$. The extension to higher-dimensional settings is straightforward in standard texts [106].

Due to the algorithm's simplicity and effectiveness, there has been considerable interest over several decades in generalizing Newton's method beyond the smooth or even finite-dimensional setting. We refer the interested reader to the monographs [49, 38, 61, 70, 125]. The setting we consider is that of *generalized equations* in Euclidean space of the form

$$0 \in F(u) + \Psi(u), \tag{4.1.1}$$

where $F : \mathbf{U} \to \mathbf{V}$ is smooth and $\Psi : \mathbf{U} \rightrightarrows \mathbf{V}$ is set-valued. One avenue to generalize Newton's method to this setting is the Josephy-Newton iteration [63] defined by

$$u \leftarrow \text{ solution } \hat{u} \text{ of } 0 \in F(u) + DF(u)(\hat{u} - u) + \Psi(\hat{u}). \tag{4.1.2}$$

Central to the analysis of algorithms of this type is metric regularity [114] of the mapping $\Phi := F + \Psi$, a generalization of invertibility, which allows us to stably bound the distance to the solution set $\Phi^{-1}(0)$ by some multiple of the *residual* function $u \mapsto \operatorname{dist}(0, \Phi(u))$. The monograph [38] is instructive as a modern reference on various notions of metric regularity with connections to implicit function theorems and applications.

Definition 4.1.1. Suppose $\bar{v} \in \Phi(\bar{u})$. We say that Φ is *metrically regular* at $\bar{u}$ for value $\bar{v}$ if there exists some $K > 0$ such that

$$\operatorname{dist}(u, \Phi^{-1}(v)) \leq K \operatorname{dist}(v, \Phi(u))$$

for all (u, v) near $(\bar{u}, \bar{v})$. If, in addition, $\operatorname{gph} \Phi^{-1}$ locally agrees with the graph of a single-valued mapping around $(\bar{y}, \bar{x})$, we say Φ is *strongly metrically regular*.

Note that in the strongly metrically regular case, Φ^{-1} is locally single-valued and Lipschitz with the same constant K ([38, Proposition 3G.1]). For example, a linear map $A : \mathbf{U} \to \mathbf{V}$ is metrically regular at 0 if and only if it is surjective, and strongly metrically regular when it is invertible. In this setting, local superlinear convergence rates of generalized Newton methods can be established [8, 51, 38, 62].

Some literature instead works with a slightly different regularity assumption originating in [117] and named in [39], which generalizes injectivity of mappings.

Definition 4.1.2. Suppose $\bar{v} \in \Phi(\bar{u})$. We say that Φ is *strongly metrically subregular* at $\bar{u}$ for value $\bar{v}$ if there exists some $K > 0$ such that

$$|u - \bar{u}| \leq K \operatorname{dist}(\bar{v}, \Phi(u))$$

for all u near $\bar{u}$.

In particular, strong metric subregularity implies $\bar{u}$ must be an isolated solution of the generalized equation, and linear $A : \mathbf{U} \to \mathbf{V}$ is strongly metrically subregular at 0 when A is injective. Under strong metric subregularity, superlinear convergence for generalized Newton and inexact quasi-Newton algorithms have been established [38, 28, 16]. In the main scenarios we consider in this chapter however, these varying metric regularity definitions are all in fact equivalent. One limitation of the iteration (4.1.2) is that the set-valued part of the equation is left untreated. The recent work [55] incorporates a "linearization" of Ψ using sophisticated graphical differentiation ideas. In this chapter, we adopt a simpler and more geometrically motivated approach.

When Ψ is the normal cone operator N_Q of some set $Q \subset \mathbf{U}$, *active-set* approaches from standard texts [49, 106] offer a useful foundation upon which to build practical algorithms. The active-set idea also extends to more contemporary optimization settings [83, 90, 73]. In a first phase, these methods *identify*, either explicitly or implicitly, some lower-dimensional "active manifold" defined by some constraints. After this identification, accelerated local convergence can be achieved, typically as a result of a linearization argument.

With roots in [130, 66, 19, 20, 22, 47, 50, 52], the theory of *partial smoothness* [79] frames active sets in a broad setting. [42] thoroughly explores the relationship with identifiability, and [80] recently extended partial smoothness to set-valued operators. Motivated by these recent works and the classical linearization interpretation of Newton's method, we develop an intuitive geometric framework for active-set Newton algorithms for generalized equations. The forthcoming manuscript [85] contains many of the major results.

4.2 A Newton method for intersecting manifolds

We begin in the general setting of two manifolds $\mathcal{X}$ and $\mathcal{Y}$ intersecting transversally, and study the intersection when one manifold is linearized. The resulting "semi-linearized" algorithm generalizes the core geometric motivation of Newton's method.

Definition 4.2.1. Let $\mathcal{X}, \mathcal{Y}$ be manifolds around $z \in \mathbf{E}$. We say that $\mathcal{X}$ and $\mathcal{Y}$ intersect *transversally* at z, or that z is a *transversal* point in $\mathcal{X} \cap \mathcal{Y}$, if

$$N_{\mathcal{X}}(z) \cap N_{\mathcal{Y}}(z) = \{0\}. \tag{4.2.1}$$

It is well known that if $\mathcal{X}, \mathcal{Y} \subset \mathbf{E}$ are manifolds intersecting transversally at z, the intersection $\mathcal{X} \cap \mathcal{Y}$ is also a manifold around z, with dimension

$$\dim(\mathcal{X} \cap \mathcal{Y}) = \dim \mathcal{X} + \dim \mathcal{Y} - \dim \mathbf{E}.$$

Proposition 4.2.1. *The following are equivalent for smooth manifolds $\mathcal{X}, \mathcal{Y}$ around $z \in \mathbf{E}$.*

(i) *$\mathcal{X}$ and $\mathcal{Y}$ intersect transversally at z.*

(ii) *There exists a local parametrization $H : \mathbf{R}^k \to \mathbf{E}$ of $\mathcal{X}$ around z and a defining map $G : \mathbf{E} \to \mathbf{R}^m$ of $\mathcal{Y}$ such that the derivative*

$$D(G \circ H)(0) : \mathbf{R}^k \to \mathbf{R}^m : w \mapsto DG(z)DH(0)w$$

is surjective.

In addition, z is isolated in $\mathcal{X} \cap \mathcal{Y}$ if and only if $k = m$ (and so the derivative is invertible).

Proof. First, assume (i) holds. Since $\mathcal{X}$, $\mathcal{Y}$ are manifolds around z, there exists a local parametrization $H : \mathbf{R}^k \to \mathbf{E}$ of $\mathcal{X}$ and a defining map $G : \mathbf{E} \to \mathbf{R}^m$

for $\mathcal{Y}$. Suppose $DH(0)^* DG(z)^* w = 0$ for some $w \in \mathbf{R}^k$. (Due to our normal rather than tangential definition of transversality, it will be more convenient to prove that the adjoint is injective.) So $DG(z)^* w \in \text{Null}(DH(0)^*) = N_{\mathcal{X}}(z)$. Since $N_{\mathcal{Y}}(z) = \text{Range}(DG(z)^*)$, by transversality $DG(z)^* w$ must be 0. But $DG(z)$ is surjective, so $w = 0$ and $D(G \circ H)(0)$ is injective. Furthermore if z is isolated in $\mathcal{X} \cap \mathcal{Y}$, near z we have that

$$\begin{aligned} 0 &= \dim(\mathcal{X} \cap \mathcal{Y}) \\ &= \dim \mathcal{X} + \dim \mathcal{Y} - \dim \mathbf{E} \\ &= k + (\dim \mathbf{E} - m) - \dim \mathbf{E} \\ &= k - m. \end{aligned}$$

Now assume (ii) holds and let $v \in N_X(z) \cap N_Y(z)$. So $DH(0)^* v = 0$ and $v = DG(z)^* w$ for some $w \in \mathbf{R}^k$. Then $DH(0)^* DG(z)^* w = 0$, which means that $w = 0$ since $D(G \circ H)(0)$ is surjective. But then $v = 0$, so transversality holds.

To see that z is an isolated intersection point when the derivative is invertible, observe that $u \in \mathcal{X} \cap \mathcal{Y}$ can be characterized locally by solutions of

$$G(H(w)) = 0.$$

By the inverse function theorem, $G \circ H$ is a bijection around 0, so $z = H(0)$ is isolated in $\mathcal{X} \cap \mathcal{Y}$. □

Theorem 4.2.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be $C^{(2)}$-manifolds around isolated transversal point $z \in \mathbf{E}$. There exists a neighbourhood U of z and a unique $C^{(1)}$ function $y : \mathcal{X} \cap U \to \mathcal{Y} \cap U$ satisfying $y(x) - x \in T_{\mathcal{X}}(x)$ with*

$$y(x) - z = O(|x - z|^2)$$

Proof. Let $H : \mathbf{R}^k \to \mathbf{E}$ and $G : \mathbf{E} \to \mathbf{R}^k$ be a local parametrization and defining map of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Thus points x near z have the form $x = H(w)$ for small $w \in \mathbf{R}^k$, and we seek a point $y(x) = H(w) + DH(w)v$ satisfying

$$G(H(w) + DH(w)v) = 0$$

for some small $v \in \mathbf{R}^k$. Denote the function on the left hand side as $F(w, v)$, and notice that $F(0, 0) = 0$, and the derivative with respect to v when $w = 0$ is $D_v F(0, v) = D_v G(H(0) + DH(0)v) = DG(H(0) + DH(0)v)DH(0)$. In particular $D_v F(0, 0) = DG(H(0))DH(0)$, which by transversality is invertible. By the implicit function theorem, there exists a unique $C^{(1)}$ function $v(w)$ defined for small w and satisfying $F(w, v(w)) = 0$ and $v(0) = 0$.

Now, H and G are $C^{(2)}$, hence in particular locally Lipschitz and we have that

$$\begin{aligned} G(H(w + v(w))) &= G(H(w) + DH(w)v(w) + O(|v(w)|^2)) \\ &= G(H(w) + DH(w)v(w)) + O(|v(w)|^2) \\ &= O(|v(w)|^2) \end{aligned}$$

as $w \to 0$, which implies $w + v(w) = O(|v(w)|^2)$ since $G(H(0)) = 0$. Hence

$$\begin{aligned} y(x) - z &= H(w) + DH(w)v(w) - H(0) \\ &= H(w + v(w)) - H(0) + O(|v(w)|^2) \\ &= O(|w + v(w)|) + O(|v(w)|^2) \\ &= O(|v(w)|^2) \\ &= O(|w|^2) \\ &= O(|H(w) - H(0)|^2) \\ &= O(|x - z|^2). \end{aligned}$$

□

If there was a way to restore the point $y(x)$ to the manifold $\mathcal{X}$ through some Lipschitz "restoration map," repeating this linearization process would generate a sequence converging quadratically to z. The following result forms the foundation of the Newton algorithms we consider in this work.

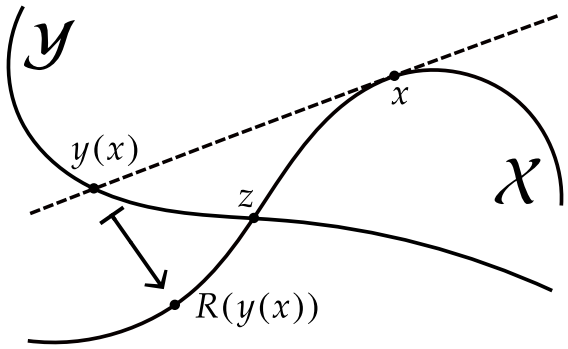


Figure 4.1: Semi-linearized Newton method.

Corollary 4.2.1. *Consider the setting of Theorem 4.2.1, and suppose there exists a Lipschitz map $R : \mathcal{Y} \to \mathcal{X}$ such that $R(z) = z$. Then given a starting point $x \in \mathcal{X}$ sufficiently close to z, the iteration*

$$x \leftarrow R(y(x)) \tag{4.2.2}$$

converges quadratically to z.

Proof. By Theorem 4.2.1, we have that for x near z,

$$|R(y(x)) - z| = |R(y(x)) - R(z)| = O(|y(x) - z|) = O(|x - z|^2).$$

□

An obvious candidate for R is the projection mapping $\mathrm{Proj}_{\mathcal{X}}$, which is well-defined and locally Lipschitz around z. But in applications, such a projection may be complex and difficult to perform, and does not provide much algorithmic

insight. We end this section by showing that when there exists a constant rank map P from $\mathcal{X}$ to $\mathcal{Y}$, it is always possible to construct a restoration map consisting of decoupled updates onto the image and preimage of P. Such a construction turns out to be a more natural choice for R, and will be the foundation of later active-set methods.

Theorem 4.2.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be $C^{(2)}$-manifolds around $z \in \mathbf{E}$. Suppose there exists a $C^{(2)}$-smooth map $P : \mathcal{X} \to \mathcal{Y}$ that is constant rank near $z = P(z)$. Then for any sufficiently small neighbourhood U of z, the following holds:*

(i) *$\mathcal{M} = P(\mathcal{X} \cap U)$ is a $C^{(2)}$-manifold of dimension equal to the rank of P at z.*

(ii) *There exists a $C^{(1)}$ map $S : \mathcal{Y} \cap U \to \mathcal{M}$ such that $S(z) = z$. A particular choice is the projection map $\mathrm{Proj}_{\mathcal{M}}$.*

(iii) *There exists a $C^{(1)}$ map $Q : \mathcal{M} \cap U \to \mathcal{X}$ such that $Q(z) = z$ and is a left inverse for P:*

$$P(Q(u)) = u \quad \text{for all } u \in \mathcal{M} \cap U.$$

A particular choice is the projection onto the preimage of u:

$$Q(u) = \mathrm{Proj}_{P^{-1}(u)}(u).$$

Proof. In the framework of the constant rank theorem, let $\phi : \mathbf{R}^k \times \mathbf{R}^{n-k} \to \mathbf{E}$ and $\psi : \mathbf{R}^k \times \mathbf{R}^{m-k} \to \mathbf{E}$ be local parametrizations around z of $\mathcal{X}$ and $\mathcal{Y}$ respectively. (i) then follows immediately, as $w \mapsto \psi(w, 0)$ is clearly a local parametrization of $\mathcal{M}$.

(ii) follows from the fact that for a $C^{(2)}$-manifold around z, the projection onto the manifold is well-defined and $C^{(1)}$ around z. (e.g., [119, see 13.38]).

To prove (iii), consider local coordinates w for some $u \in \mathcal{M}$. By the constant rank theorem, the preimage $P^{-1}(u)$ has a local parametrization $v \mapsto \phi(w, v)$, and is thus a $C^{(2)}$-manifold around $\phi(w, 0)$. So for w sufficiently small, the projection of u onto this manifold is well-defined with local coordinates given by the global minimizer of $\min_v |\phi(w, v) - \psi(w, 0)|^2$. This minimizer must satisfy the first-order necessary condition

$$D_v\phi(w, v)^*(\phi(w, v) - \psi(w, 0)) = 0,$$

which is $C^{(1)}$-smooth as a function of (w, v), and clearly satisfied when $(w, v) = (0, 0)$. The derivative with respect to v of the left hand side when $(w, v) = (0, 0)$ is $D_v\phi(0, 0)^* D_v\phi(0, 0)$, which is invertible since $D\phi(0, 0)$ is injective. Therefore by the implicit function theorem, for small w the first-order condition has a unique small solution $v(w)$ that is $C^{(1)}$-smooth as a function of w, which implies that $\phi(w, v(w))$ must be exactly $\mathrm{Proj}_{P^{-1}(u)}(u)$. □

Example (Linear maps). Consider a rank k linear map A between Euclidean spaces $\mathcal{X} = \mathbf{E}$ and $\mathcal{Y} = \mathbf{F}$. The manifold $\mathcal{M}$ is then simply the range of A, and Proj_M the orthogonal projection onto this subspace. Given $y \in \mathrm{Range}(A)$, we can define the left inverse $Q(y) = A^\dagger y$, where $A^\dagger$ is the *pseudo-inverse* of A satisfying $AA^\dagger y = y$. It is well known (see e.g., [56, Section 5.5]) that $Q(y)$ then gives the minimum norm solution of $Ax = y$.

4.3 Partly smooth generalized equations

Our aim is to apply the results of the previous section to generalized equations of the form

$$v \in \Phi(u) \tag{4.3.1}$$

involving some partly smooth (Definition 2.5.6) set-valued mapping $\Phi : \mathbf{U} \rightrightarrows \mathbf{V}$. In particular, we assume that the graph

$$\operatorname{gph} \Phi = \{(u, v) \in \mathbf{U} \times \mathbf{V} : v \in \Phi(u)\}.$$

is a smooth manifold around $(\bar{u}, \bar{v}) \in \mathbf{U} \times \mathbf{V}$. Given $\bar{v}$, we cast the problem of solving $\bar{v} \in \Phi(u)$ as a manifold intersection problem, and apply the results of the previous section with $\mathcal{X} = \operatorname{gph} \Phi$ and $\mathcal{Y} = \mathbf{U} \times \{\bar{v}\}$. In this setting it will be convenient to adopt the language of *graphical differentiation* (Definition 2.3.1). In this language, the Newton step of Theorem 4.2.1 can be performed by solving

$$\bar{v} - v \in D\Phi(u|v)(w - u) \tag{4.3.2}$$

for w. Since $T_{\operatorname{gph} \Phi}(u, v)$ and $\mathbf{U} \times \{\bar{v}\}$ are both affine, if $(\bar{u}, \bar{v})$ is an isolated solution, the transversality condition (4.2.1) guarantees that w is uniquely determined. This is in fact equivalent to metric regularity of Φ, which we can conveniently characterize via the coderivative.

Proposition 4.3.1. *The following are equivalent when Φ is a smooth manifold around $(\bar{u}, \bar{v})$.*

(i) $\operatorname{gph} \Phi$ *intersects* $\mathbf{U} \times \{\bar{v}\}$ *transversally at* $(\bar{u}, \bar{v})$.

(ii) $0 \in D^*\Phi(\bar{u}|\bar{v})(v) \Rightarrow v = 0$.

(iii) Φ *is metrically regular at* $\bar{u}$ *for value* $\bar{v}$.

Proof. Assume transversality, i.e.,

$$N_{\operatorname{gph} \Phi}(\bar{u}, \bar{v}) \cap N_{\mathbf{U} \times \{\bar{v}\}} = \{0\}.$$

Noting that $N_{\mathbf{U} \times \{\bar{v}\}}(\bar{u}, \bar{v}) = \{0\} \times \mathbf{V}$, we have that

$$(0, v) \in N_{\operatorname{gph} \Phi}(\bar{u}, \bar{v}) \Rightarrow v = 0,$$

which is by definition

$$0 \in D^*\Phi(\bar{u}|\bar{v})(-v) \Rightarrow v = 0.$$

Since v was arbitrary, (ii) follows. The reverse implication proceeds identically. The equivalence (ii) $\Leftrightarrow$ (iii) is [38, Theorem 4C.2]. □

Example (Nonlinear equations). Consider the system of equations $F(u) = 0$ for single-valued $C^{(2)}$-smooth $F : \mathbf{R}^n \to \mathbf{R}^n$. Transversality then amounts to the Jacobian ∇F being invertible at solution $\bar{u}$. Given $(u, F(u))$ near $(\bar{u}, 0)$, the Newton step (4.3.2) becomes exactly the classical Newton iteration

$$w = u - \nabla F(u)^{-1} F(u).$$

If we choose the "natural" restoration map $R : (w, 0) \mapsto (w, F(w))$, then Theorem 4.2.1 recovers the local quadratic convergence of Newton's method.

Example (Monotone operators). Consider the equation $0 \in \Phi(u)$ for *maximally monotone* $\Phi : \mathbf{U} \rightrightarrows \mathbf{U}$ [116]. In particular this case includes the optimality condition $0 \in \partial f(u)$ for proper closed convex function $f : \mathbf{U} \to \bar{\mathbf{R}}$. Following the Newton step

$$w = u + D\Phi(u|v)^{-1}(-v),$$

a natural choice for the restoration map makes use of the fact that for any $\lambda > 0$, the *resolvent*

$$J_{\lambda\Phi} = (I + \lambda\Phi)^{-1}$$

is single-valued and nonexpansive. A simple calculation shows that $J_{\lambda\Phi}(\bar{u}) = \bar{u}$ and that

$$\Big(J_{\lambda\Phi}(u), \frac{1}{\lambda}\big(u - J_{\lambda\Phi}(u)\big)\Big) \in \operatorname{gph}\Phi \tag{4.3.3}$$

for all $u \in \mathbf{U}$. Hence we can define a restoration map R as the function that maps $(u, 0)$ to the left hand side of (4.3.3).

Even for monotone Φ, calculating the resolvent may be as difficult as inverting Φ itself. To work around this, one approach may be to decompose Φ into simpler mappings that can be handled independently, in the vein of *splitting methods*. We explore some preliminary algorithms in this direction in Section 4.6. Here, using the theory of partial smoothness, we take a different direction that does not require monotonicity.

Theorem 4.2.2 describes a decoupled approach to the restoration map R, assuming the existence of some constant rank map between the manifolds of interest. Recall that if Φ is partly smooth at $\bar{u}$ for $\bar{v}$, the projection map $(u, v) \mapsto u$ is constant rank around $(\bar{u}, \bar{v})$, and its image is some *active manifold* $\mathcal{M}$. By specializing to partly smooth generalized equations, the structure of the decoupled restoration map becomes clear. A primal update that restores the variable u to $\mathcal{M}$ is followed by a dual update that fixes u and chooses an appropriate $v \in \Phi(u)$.

Theorem 4.3.1. *Let $\bar{u}$ be an isolated solution to $\bar{v} \in \Phi(u)$, and suppose $\Phi : \mathbf{U} \to \mathbf{V}$ is $C^{(2)}$-partly smooth at $\bar{u}$ for $\bar{v}$ with active manifold $\mathcal{M} \subset \mathbf{U}$. Then assuming the transversality condition*

$$0 \in D^*\Phi(\bar{u}|\bar{v})(v) \Rightarrow v = 0,$$

for any sufficiently small neighbourhood U of $\bar{u}$ the following holds.

(i) *There exists a $C^{(1)}$ map $S : U \to \mathcal{M}$ such that $S(\bar{u}) = \bar{u}$. A particular choice is the projection map* $\mathrm{Proj}_{\mathcal{M}}$.

(ii) *There exists a $C^{(1)}$ map $Q : \mathcal{M} \cap U \to \mathbf{V}$ such that $Q(\bar{u}) = \bar{v}$ and*

$$Q(u) \in \Phi(u) \quad \textit{for all } u \in \mathcal{M} \cap U.$$

A particular choice is

$$Q(u) = \mathrm{Proj}_{\Phi(u)}(\bar{v}).$$

(iii) *For all* $(u, v) \in \operatorname{gph} \Phi$ *near* $(\bar{u}, \bar{v})$*, the linearized equation*

$$\bar{v} - v \in D\Phi(u|v)(w - u)$$

has a unique solution $w(u, v)$*.*

Moreover, given any starting point $(u, v) \in \operatorname{gph} \Phi$ *sufficiently close to* $(\bar{u}, \bar{v})$*, the iteration*

$$u \leftarrow S(w(u, v)), \qquad v \leftarrow Q(u)$$

converges quadratically to $(\bar{u}, \bar{v})$*.*

Proof. Apply Corollary 4.2.1 and Theorem 4.2.2 with

$$\mathcal{X} = \operatorname{gph} \Phi, \qquad \mathcal{Y} = \mathbf{U} \times \{\bar{v}\},$$
$$z = (\bar{u}, \bar{v}), \qquad P(u, v) = (u, \bar{v}).$$

□

4.4 Identification and smooth reductions

While the algorithm described in Theorem 4.3.1 is elegant and conceptually appealing, several issues remain a barrier to a practical implementation. As in any purely local algorithm, the question of how to initialize close to a solution of interest is paramount. The algorithms also involves the potentially complex operations of projecting onto the active manifold $\mathcal{M}$ and images $\Phi(u)$, as well as inverting the graphical derivative $D\Phi(u|v)$. The aim of this section is to address these issues for a broad class of interesting applications.

Consider the generalized equation

$$0 \in \Phi(u) \tag{4.4.1}$$

where $\Phi : \mathbf{U} \rightrightarrows \mathbf{U}$ is partly smooth at $\bar{u}$ for 0 with active manifold $\mathcal{M} \subset \mathbf{U}$. We focus on the case where Φ *locally reduces* to an operator of the form $F + N_{\mathcal{M}}$, in the sense that

$$\operatorname{gph} \Phi = \operatorname{gph}(F + N_{\mathcal{M}}) \quad \text{near } (\bar{u}, 0) \tag{4.4.2}$$

for some smooth function $F : \mathbf{U} \to \mathbf{U}$. We will show that this assumption, while stringent, holds for a variety of interesting cases.

Partial smoothness is closely linked to the notion of *identifiable sets*. Indeed, the definition of a partly smooth operator immediately yields the identification property

$$v_k \in \Phi(u_k) \text{ and } (u_k, v_k) \to (\bar{u}, 0) \implies u_k \in \mathcal{M} \text{ for all large } k, \tag{4.4.3}$$

which is remarkably broad and powerful. Given any algorithm producing a sequence u_k, under a metric regularity assumption it is natural to seek a sequence $v_k \in \Phi(u_k)$ converging to 0 as a guarantee that u_k converges to $\bar{u}$. In this case (4.4.3) says that u_k must also identify $\mathcal{M}$.

The identification property and smooth reduction highlight the essence of active-set strategies. Some iterative algorithm that identifies the active manifold can be employed in a "global phase," which is followed by a "local phase" where the semi-linearized Newton method of Theorem 4.3.1 is performed (implicitly or explicitly) on the simpler mapping $F + N_{\mathcal{M}}$.

Example (Variational inequalities). Following the terminology of [49], a *variational inequality* $\mathrm{VI}(Q, F)$ is a generalized equation of the form

$$0 \in F(u) + N_Q(u)$$

for $C^{(2)}$-smooth $F : \mathbf{U} \to \mathbf{U}$ and closed set $Q \subseteq \mathbf{U}$.

Let $\bar{u}$ be an isolated solution to VI(Q, F), and assume that Q is $C^{(3)}$-partly smooth at $\bar{u}$ for $-F(\bar{u})$ along with the standard nondegeneracy condition

$$-F(\bar{u}) \in \operatorname{ri} \hat{N}_Q(\bar{u}).$$

By Theorem 2.5.1, after applying a sum rule we have that $F + N_Q$ is $C^{(2)}$-partly smooth at $\bar{u}$ for 0 and that

$$\operatorname{gph} N_Q = \operatorname{gph} N_{\mathcal{M}} \quad \text{near } (\bar{u}, -F(\bar{u})). \tag{4.4.4}$$

To illustrate identification, for simplicity suppose that Q is convex. Consider the mapping T defined by

$$T(u) = \operatorname{Proj}_Q(u - F(u)).$$

The fact that $\bar{u}$ solves VI(Q, F) if and only if $T(\bar{u}) = \bar{u}$ motivates the basic fixed point iteration

$$u \leftarrow T(u).$$

Under reasonable conditions (see e.g., [49, Theorem 12.1.2]), this iteration yields a sequence u_k converging to a solution $\bar{u}$. By continuity,

$$T(u_k) \to \bar{u} \quad \text{and} \quad u_k - T(u_k) - F(u_k) \to -F(\bar{u}),$$

and hence

$$u_k - T(u_k) - F(u_k) \in N_Q(T(u_k)).$$

Therefore by the reduction (4.4.4) we have that $u_k \in \mathcal{M}$ for all k sufficiently large. More sophisticated global algorithms for VI(Q, F) aim to drive the natural *residual* $r_k = u_k - T(u_k)$ to zero by way of a *merit function*. A similar calculation shows that $u_k + r_k \in \mathcal{M}$ eventually.

Example (Minimizing partly smooth functions). Consider the optimization problem

$$\min_u f(u)$$

for a closed function $f : \mathbf{U} \to \bar{\mathbf{R}}$. Suppose that $\bar{u}$ is a nondegenerate critical point:

$$0 \in \operatorname{ri} \hat{\partial} f(\bar{u}),$$

and that f is partly smooth at $\bar{u}$ for 0 relative to manifold $\mathcal{M}$. By Theorem 2.5.1 the subdifferential satisfies

$$\operatorname{gph} \partial f = \operatorname{gph}(\nabla \bar{f} + N_{\mathcal{M}}) \quad \text{near } (\bar{u}, 0),$$

where $\bar{f}$ is a smooth function agreeing with f on $\mathcal{M}$. Therefore Theorem 4.3.1 can be employed in a local phase following any global algorithm that generates small subgradients.

Returning to the general case, we next discuss how to compute with the operator $\Phi = F + N_{\mathcal{M}}$. It will be convenient to fix $G : \mathbf{U} \to \mathbf{R}^m$ as a defining map for $\mathcal{M}$ and define the function

$$H : \mathbf{U} \times \mathbf{R}^m \to \mathbf{U} : (u, \lambda) \mapsto DG(u)^* \lambda.$$

Note that since $DG(u)$ is surjective, $H(u, \cdot)$ is a local parametrization of $N_{\mathcal{M}}(u)$, and we will refer to λ as the local coordinates of some $y \in N_{\mathcal{M}}(u)$.

In the setting of Theorem 4.3.1, we first address the restoration map R. The exact manifold projection $\operatorname{Proj}_{\mathcal{M}}$ is often intractable and unnecessary in practice, but required for our theoretical results. In the case of a variational inequality $\operatorname{VI}(Q, F)$ for partly smooth Q, we could instead use Proj_Q, which agrees with

$\mathrm{Proj}_{\mathcal{M}}$ locally. The graph restoration $\mathrm{Proj}_{F(u)+N_{\mathcal{M}}(u)}(0)$ can be computed by solving

$$\min_{\lambda \in \mathbf{R}^m} \left| F(u) + \nabla G(u)^* \lambda \right|,$$

a simple least-squares problem.

Turning now to the Newton step, by the sum rule for graphical derivatives (4.3.2) becomes

$$-v \in \nabla F(u)(w-u) + DN_{\mathcal{M}}(u|v - F(u))(w-u), \tag{4.4.5}$$

with transversality becoming

$$-\nabla F(\bar{u})^* v \in D^* N_{\mathcal{M}}(\bar{u}|-F(\bar{u}))(v) \Rightarrow v = 0. \tag{4.4.6}$$

Since $N_{\mathcal{M}}$ is the subdifferential of the indicator function $\delta_{\mathcal{M}}$, $DN_{\mathcal{M}}$ and $D^*N_{\mathcal{M}}$ are second-order objects. The coderivative in particular has played an important role in generalized second-order theory, with the Mordukhovich *generalized Hessian* [99] of $\delta_{\mathcal{M}}$ defined as

$$\partial^2 \delta_{\mathcal{M}} = D^*(\partial \delta_{\mathcal{M}}) = D^* N_{\mathcal{M}}.$$

It is well known that Hessian of a $C^{(2)}$-smooth $f : \mathbf{U} \to \mathbf{R}$ function is symmetric, i.e.,

$$D(\nabla f)(u) = \nabla^2 f(u) = D^*(\nabla f)(u).$$

For general sets $\mathcal{M}$, there does not exist a similar relationship between the coderivative and graphical derivative. However for smooth manifolds, by a result of [86], $DN_{\mathcal{M}}$ and $D^*N_{\mathcal{M}}$ are in fact equivalent, and admit simple representations.

Theorem 4.4.1. *Consider $N_{\mathcal{M}} : \mathbf{U} \rightrightarrows \mathbf{U}$ for a $C^{(r)}$-smooth ($r \geq 2$) manifold $\mathcal{M}$ of codimension m around $u \in \mathbf{U}$. For any normal vector $y \in N_{\mathcal{M}}(u)$ with local coordinates*

$\lambda \in \mathbf{R}^m$, *the graph of* $N_{\mathcal{M}}$ *is a* $C^{(r-1)}$*-manifold of dimension* $\dim \mathbf{U}$ *around* (u, y) *with derivatives*

$$DN_{\mathcal{M}}(u|y)(w) = D^*N_{\mathcal{M}}(u|y)(w) = \begin{cases} D_u H(u,\lambda)w + N_{\mathcal{M}}(u) & w \in T_{\mathcal{M}}(u), \\ \varnothing & w \notin T_{\mathcal{M}}(u). \end{cases}$$

Proof. Let $\phi : \mathbf{R}^n \to \mathbf{U}$ be a local parametrization of $\mathcal{M}$ (and hence $n = \dim \mathbf{U} - m$). Define the map

$$P : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{U} \times \mathbf{U} : (\alpha, \lambda) \mapsto \left(\phi(\alpha), DG(\phi(\alpha))^*\lambda\right),$$

and its derivative in matrix block form

$$DP(\alpha,\lambda) = \begin{pmatrix} D\phi(\alpha) & 0 \\ \sum_{i=1}^m \lambda_i D_\alpha(DG_i(\phi(\alpha))) & DG(\phi(\alpha))^* \end{pmatrix}.$$

Since $DP(\alpha, \lambda)$ is clearly full rank, P is a local parametrization of $\operatorname{gph} N_{\mathcal{M}}$, which in particular implies that

$$\dim(\operatorname{gph} N_{\mathcal{M}}) = n + m = \dim \mathbf{U}.$$

To compute the tangent and normal spaces to $\operatorname{gph} N_{\mathcal{M}}$, we apply [86, Theorem 2.8], whose proof we reproduce in our notation. Let $(u, y) \in \operatorname{gph} N_{\mathcal{M}}$ with local coordinates (α, λ). Then we have that

$$\begin{aligned} (w,z) \in T_{\operatorname{gph} N_{\mathcal{M}}}(u,y) &\Leftrightarrow (w,z) \in \operatorname{Range}(DP(\alpha,\lambda)) \\ &\Leftrightarrow \begin{cases} w \in \operatorname{Range}(D\phi(\alpha)), \\ z \in \left(\sum_{i=1}^m \lambda_i D_u(DG_i(u))\right)w + \operatorname{Range}(DG(\phi(\alpha))^*) \end{cases} \\ &\Leftrightarrow \begin{cases} w \in T_{\mathcal{M}}(u), \\ z - D_u H(u,\lambda)w \in N_{\mathcal{M}}(u). \end{cases} \end{aligned}$$

Now use the fact that for any linear map A and subspace S,

$$\{x : Ax \in S\}^{\perp} = A^* S^{\perp}.$$

By applying this with

$$A = \begin{pmatrix} I & 0 \\ -D_u H(u,\lambda) & I \end{pmatrix}, \qquad S = \big\{(a,b) : a \in T_{\mathcal{M}}(u), b \in N_{\mathcal{M}}(u)\big\},$$

since G is $C^{(2)}$ we have that

$$(w,z) \in N_{\operatorname{gph} N_{\mathcal{M}}}(u,y) \Leftrightarrow \begin{cases} w + D_u H(u,\lambda) z \in N_{\mathcal{M}}(u), \\ z \in T_{\mathcal{M}}(u). \end{cases}$$

The result then follows from the definition of the derivative and coderivative. □

When specialized to partly smooth functions, an elegant consequence emerges. Informally, generalized second derivatives of partly smooth functions are symmetric.

Corollary 4.4.1. *Consider a subdifferentially continuous function $f : \mathbf{U} \to \bar{\mathbf{R}}$ with subgradient $\bar{v} \in \operatorname{ri} \partial f(\bar{u})$. If f is $C^{(r)}$ partly smooth (for $r \geq 2$) at $\bar{u}$ for $\bar{v}$ relative to manifold $\mathcal{M}$, then*

$$D(\partial f)(\bar{u}|\bar{v}) = D^*(\partial f)(\bar{u}|\bar{v}) = \partial^2 f(\bar{u}|\bar{v}) = \nabla^2 \bar{f}(\bar{u}) + DN_{\mathcal{M}}(\bar{u}|\bar{v} - \nabla \bar{f}(\bar{u}))$$

where $\bar{f} : \mathbf{U} \to \mathbf{R}$ is a $C^{(r)}$-smooth function agreeing with f on $\mathcal{M}$.

Proof. By Theorem 2.5.1, ∂f admits the local representation

$$\operatorname{gph} \partial f = \operatorname{gph}(\nabla \bar{f} + N_{\mathcal{M}}) \quad \text{near } (\bar{u}, \bar{v}),$$

so the result follows by routine derivative-coderivative calculus and applying Theorem 4.4.1. □

Now, returning to the setting of $\Phi = F + N_{\mathcal{M}}$, given some $(u, v) \in \operatorname{gph} \Phi$ near $(\bar{u}, 0)$ we can write $v = L(u, \lambda)$ for local coordinates λ in terms the *Lagrangian* function

$$L(u, \lambda) = F(u) + H(u, \lambda).$$

Then the solution w of the Newton step (4.4.5) can be computed by solving the linear system

$$\begin{cases} L(u, \lambda) + D_u L(u, \lambda) d \in N_{\mathcal{M}}(u) \\ d \in T_{\mathcal{M}}(u), \end{cases} \tag{4.4.7}$$

and setting $w = u + d$.

The transversality condition (4.4.6) simply amounts to invertibility of the system (4.4.7) around a solution $\bar{u}$, or equivalently, to the invertibility of $D_u L$ projected onto the tangent space $T_{\mathcal{M}}(\bar{u})$.

Corollary 4.4.2. *Let $0 \in F(\bar{u}) + N_{\mathcal{M}}(\bar{u})$, and $\bar{\lambda}$ be the local coordinates of $-F(\bar{u})$ in $N_{\mathcal{M}}(\bar{u})$. Let $Z : \mathbf{R}^{n-m} \to T_{\mathcal{M}}(\bar{u})$ be an injective linear map parameterizing the tangent space $T_{\mathcal{M}}(\bar{u})$. Then the following are equivalent.*

(i) *$F + N_{\mathcal{M}}$ is metrically regular at $\bar{u}$ for value 0.*

(ii) *$\operatorname{gph}(F + N_{\mathcal{M}})$ intersects $\mathbf{U} \times \{0\}$ transversally at the isolated point $(\bar{u}, 0)$.*

(iii) *$-DF(\bar{u})^* v \in D^* N_{\mathcal{M}}(\bar{u} | -F(\bar{u}))(v) \Rightarrow v = 0$.*

(iv) *$v \in T_{\mathcal{M}}(\bar{u})$ and $D_u L(\bar{u}, \bar{\lambda}) v \in N_{\mathcal{M}}(\bar{u}) \Rightarrow v = 0$.*

(v) *The linear map $Z^* D_u L(\bar{u}, \bar{\lambda}) Z$ is invertible.*

Proof. The equivalence of (iv) and (v) follows from [9, Proposition 14.1]. Since the proof is simple, we reproduce it here for completeness.

((i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii)). Since $\mathrm{gph}(F + N_{\mathcal{M}})$ and $\mathbf{U} \times \{0\}$ are both submanifolds of dimension $\dim \mathbf{U}$ in $\mathbf{U} \times \mathbf{U}$, if they intersect transversally, the intersection is a submanifold of dimension 0, so any transversal point is automatically isolated. Thus the equivalences follow from Proposition 4.3.1 and the sum rule for coderivatives.

(iii) $\Leftrightarrow$ (iv) follows immediately from Theorem 4.4.1.

((iv) $\Leftrightarrow$ (v)). Suppose $Z^* D_u L(\bar{u}, \bar{\lambda}) Z w = 0$. Then $Zw \in \mathrm{Range}(Z) = T_{\mathcal{M}}(\bar{u})$ and $D_u L(\bar{u}, \bar{\lambda}) Z w \in \mathrm{Null}(Z^*) = N_{\mathcal{M}}(\bar{u})$. So by (iii) we have $Zw = 0$, which implies $w = 0$ since Z is injective. On the other hand, if $v \in T_{\mathcal{M}}(\bar{u})$ and $D_u L(\bar{u}, \bar{\lambda}) v \in N_{\mathcal{M}}(\bar{u})$, then $v = Zw$ for some w and $Z^* D_u L(\bar{u}, \bar{\lambda}) Z w = 0$, which implies $v = 0$ if $Z^* D_u L(\bar{u}, \bar{\lambda}) Z$ is invertible. □

Finally, we note that various notions of metric regularity of subdifferentials are in fact equivalent in the partly smooth setting.

Corollary 4.4.3. *Given a closed function $f : \mathbf{E} \to \bar{\mathbf{R}}$, the following are equivalent when ∂f is partly smooth at $\bar{u}$ for value 0.*

(i) *∂f is metrically regular at $\bar{u}$ for value 0.*

(ii) *∂f is strongly metrically regular at $\bar{u}$ for value 0.*

(iii) *∂f is strongly metrically subregular at $\bar{u}$ for value 0.*

Proof. ∂f is strongly metrically regular when ∂f^{-1} has a single-valued Lipschitz continuous localization around 0 for $\bar{u}$, so (i) $\Leftrightarrow$ (ii) follows directly from from Corollary 4.4.2 and the inverse function theorem. This mirrors an analogous result of Dontchev and Rockafellar for KKT mappings [38, Theorem 4I.2]

(ii) $\Leftrightarrow$ (iii) follows from graphical derivative characterization of strong metric subregularity ([38, Theorem 4E.1])

$$D(\partial f)(\bar{u}|0)^{-1}(0) = \{0\}.$$

Since the graphical derivative and coderivative are equivalent in this setting, the results follows from Corollary 4.4.2. □

4.5 Example: Smooth optimization and SQP

Classical theory of nonlinear programming offers connections and provides a nice illustration of the results in this chapter. Let $f : \mathcal{M} \to \mathbf{R}$ be a $C^{(3)}$ function defined on a $C^{(3)}$-manifold $\mathcal{M} \subset \mathbf{R}^n$, and consider the corresponding extended value function

$$\tilde{f} = \begin{cases} f(u) & (u \in \mathcal{M}) \\ +\infty & (u \notin \mathcal{M}) \end{cases}$$

and subdifferential

$$\partial \tilde{f}(u) = \begin{cases} \nabla f(u) + N_{\mathcal{M}}(u) & (u \in \mathcal{M}) \\ \varnothing & (u \notin \mathcal{M}). \end{cases}$$

Letting $G : \mathbf{R}^m \to \mathbf{R}^n$ be a defining map for $\mathcal{M}$, critical points $\bar{u}$ satisfy

$$-\nabla f(\bar{u}) = \sum_{i=1}^{m} \bar{\lambda}_i \nabla G_i(\bar{u})$$

for some $\lambda \in \mathbf{R}^m$. When $\bar{u}$ is a nondegenerate critical point, meaning $\bar{\lambda} > 0$, $\partial \tilde{f}$ is partly smooth at $\bar{u}$ for 0.

Suppose furthermore that $\bar{u}$ a local minimizer around which f grows quadratically, i.e., for some $\epsilon > 0$,

$$f(u) \geq f(\bar{u}) + \epsilon \, |u - \bar{u}|^2 \quad \text{for } u \in \mathcal{M} \text{ near } \bar{u}.$$

Then defining the Lagrangian $\mathcal{L}(u, \lambda) = f(u) + \sum_{i=1}^{m} \lambda_i G_i(u)$, we have the *second-order sufficient condition* that

$$\nabla^2_{uu} \mathcal{L}(\bar{u}, \bar{\lambda}) = \nabla^2 f(\bar{u}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla^2 G_i(\bar{u})$$

is positive definite on the null space of $\nabla G(\bar{u})$. In this case condition (v) of Theorem 4.4.2 holds, and transversality condition

$$-\nabla^2 f(\bar{u})v \in D^* N_{\mathcal{M}}(\bar{u}, -\nabla f(\bar{u}))(v) \Rightarrow v = 0$$

follows.

Now suppose $(u, v) \in \operatorname{gph} \partial \tilde{f}$ is close to $(\bar{u}, 0)$. Letting λ be the local coordinates of $v - \nabla f(u)$ in $N_{\mathcal{M}}(u)$, the system (4.4.7) defining the Newton step becomes

$$\begin{cases} \nabla f(u) + \nabla G(u)^* \lambda + \nabla^2_{uu} \mathcal{L}(u, \lambda) d \in N_{\mathcal{M}}(u) \\ d \in T_{\mathcal{M}}(u), \end{cases}$$

which in matrix form is

$$\begin{pmatrix} \nabla^2_{uu} \mathcal{L}(u, \lambda) & \nabla G(u)^{\mathsf{T}} \\ \nabla G(u) & 0 \end{pmatrix} \begin{pmatrix} d \\ \mu \end{pmatrix} = \begin{pmatrix} -\nabla f(u) \\ 0 \end{pmatrix}.$$

This can be immediately recognized as the first-order stationarity conditions of the quadratic program

$$\min_{d \in \mathbf{U}} \left\{ \langle \nabla f(u), d \rangle + \langle d, \nabla^2_{uu} \mathcal{L}(u, \lambda) d \rangle : \nabla G(u) d = 0 \right\},$$

a familiar subproblem in sequential quadratic programming algorithms.

At this point our algorithm diverges slightly from classical SQP methods in requiring that $u + d$ is restored to the manifold $\mathcal{M}$. In practice this is far too stringent, and SQP algorithms typically only achieve feasibility in the limit.

However, robust implementations typically include a second-order "constraint-restoration" step $e = O(|d|^2)$ of the form

$$e = -\nabla G(u)^\dagger G(u+d)$$

where $\nabla G(u)^\dagger$ is a right inverse of $\nabla G(u)$. Then $e \in N_{\mathcal{M}}(u)$ and is a step towards the manifold, so $u + d + e$ can be viewed as an approximate manifold projection. See [9, Chapter 17] or [98] for other discussions of this. Nonetheless, for our theoretical purposes we require exact feasibility, so let $w = \mathrm{Proj}_{\mathcal{M}}(u+d)$.

To perform the graph restoration step which projects 0 onto $\partial \tilde{f}(w)$, we can solve the least-squares problem

$$\min_{\lambda} \left\| \nabla f(w) + \nabla G(w)^\intercal \lambda \right\|_2^2.$$

The solution $\lambda(w)$ is given by

$$\lambda(w) = -(\nabla G(w)^\intercal)^\dagger \nabla f(w),$$

where $(\nabla G(w)^\intercal)^\dagger$ is a left inverse for $\nabla G(w)^\intercal$, the *least-squares update* of the dual variables in SQP terminology. Then

$$w \mapsto \nabla f(w) + \nabla G(w)^\intercal \lambda(w)$$

is exactly the projection we seek.

4.6 A second-order forward-backward method

As we alluded to in Section 4.3, projections are not the only way to build restoration maps. An alternative is to use the resolvent $(I + \lambda\Psi)^{-1}$ of monotone operators Ψ. Consider the generalized equation

$$0 \in \Phi(u) := F(u) + \Psi(u) \tag{4.6.1}$$

for $C^{(2)}$-smooth $F : \mathbf{U} \to \mathbf{U}$ and maximal monotone $\Psi : \mathbf{U} \rightrightarrows \mathbf{U}$. Suppose that Ψ is partly smooth with respect to $\mathcal{M} \subset \mathbf{U}$ at $\bar{u}$ for value $-F(\bar{u})$. Hence Φ is partly smooth at $\bar{u}$ for 0 with respect to the same manifold. Furthermore suppose that $\bar{u}$ is an isolated solution, and the transversality condition

$$-\nabla F(\bar{u})^* v \in D^*\Psi(\bar{u}|-F(\bar{u}))(v) \Rightarrow v = 0 \tag{4.6.2}$$

holds.

A popular method for solving 4.6.1 is *forward-backward splitting* [3, Section 26.5]. The basic steps of the algorithm consist of a "forward" step $u \mapsto u - tF(u)$ followed by an implicit "backward" step $u \mapsto (I + t\Psi)^{-1}u$. (For simplicity, we fix the step size $t > 0$ whereas in practice it may be allowed to vary at each iteration.)

Defining the forward-backward map

$$T(u) = (I + t\Psi)^{-1}(u - tF(u)),$$

it is clear that $0 \in \Phi(\bar{u})$ if and only if $\bar{u}$ is a fixed point of T. An easy manipulation shows that

$$\left(T(u), \frac{1}{t}(u - T(u)) - F(u) + F(T(u))\right) \in \operatorname{gph} \Phi.$$

Since the resolvent operator $(I+t\Psi)^{-1}$ is nonexpansive, and F is locally Lipschitz, we can define the Lipschitz restoration map

$$(u, 0) \mapsto \left(T(u), \frac{1}{t}(u - T(u)) - F(u) + F(T(u))\right) \tag{4.6.3}$$

that fixes the point $(\bar{u}, 0)$. Applying our linearization framework immediately yields the following algorithm, which is simply the regular forward-backward method, but augmented with second-order steps once the active manifold has been identified.

Algorithm 4.1: Second-order forward backward algorithm to solve partly smooth $0 \in F(u) + \Psi(u)$

Input : Starting point $(u, v) \in \operatorname{gph} \Phi$, step size $t > 0$, stopping tolerance ϵ;
while $|v| > \epsilon$ **do**
  if $u \in \mathcal{M}$ **then**
    Solve $-v \in DF(u)w + D\Psi(u|v - F(u))(w)$ for w;
    $u \leftarrow u + w$;
  end
  $v \leftarrow \frac{1}{t}(u - T(u)) - F(u) + F(T(u))$;
  $u \leftarrow T(u)$;
end

A case of particular interest arises in considering sum-composite optimization problems of the form

$$\min_{x \in \mathbf{E}} \; f(x) + g(x)$$

for smooth, convex $f : \mathbf{E} \to \mathbf{R}$ with Lipschitz gradients and closed, convex, proper $g : \mathbf{E} \to \bar{\mathbf{R}}$. Minimizers $\bar{x}$ are characterized by the stationarity condition

$$0 \in \nabla f(\bar{x}) + \partial g(\bar{x}),$$

which fits into the form of (4.6.1). In this case, the forward-backward method is usually called the *proximal-gradient* algorithm, since

$$\begin{aligned} T(x) &= \left(I + t\partial g\right)^{-1}\left(x - t\nabla f(x)\right) \\ &= \operatorname{prox}_{tg}\left(x - t\nabla f(x)\right) \\ &= \operatorname*{arg\,min}_{z \in \mathbf{E}} \left\{ f(x) + \langle \nabla f(x), z - x \rangle + g(z) + \frac{1}{2t}\,|z - x|^2 \right\} \end{aligned}$$

In this sum-composite optimization setting, the Newton step of Algorithm 4.1 can be written

$$-v \in \nabla^2 f(x)w + \partial^2 g(x|v - \nabla f(x))(w). \tag{4.6.4}$$

In many applications g is a simple partly smooth function, allowing us to calculate $\partial^2 g$ analytically, which leads to simple implementable forms of Algorithm 4.1. Suppose that $\bar{x}$ is a nondegenerate strict local minimizer, meaning

$$-\nabla f(\bar{x}) \in \mathrm{ri}(\partial g(\bar{x})),$$

and that g is partly smooth at $\bar{x}$ for value $-\nabla f(\bar{x})$ relative to the manifold $\mathcal{M} \subset \mathbf{E}$. By Theorem 2.5.1, the graph of ∂g decomposes as

$$\mathrm{gph}\, \partial g = \mathrm{gph}(\nabla \bar{g} + N_{\mathcal{M}}) \quad \text{near } (\bar{x}, -\nabla f(\bar{x})),$$

where $\bar{g}$ is some $C^{(2)}$-smooth function agreeing with g on $\mathcal{M}$. Then we can apply Theorem 4.4.1 and Corollary 4.4.1 to calculate $\partial^2 g$.

Algorithm 4.2: Second-order proximal gradient with damped linesearch

Input : Starting point $(u, v) \in \mathrm{gph}\, \Phi$, step size $t > 0$, linesearch parameters $\beta \in (0, 1)$, $r_{\min} > 0$, stopping tolerance ϵ;
while $|v| > \epsilon$ **do**
  if *try Newton step* **then**
    Solve $-v \in \nabla^2 f(x)w + \partial^2 g(x|v - \nabla f(x))(w)$ for w;
    $\bar{r} = \sup\left\{r \in \{1, \beta, \beta^2, \ldots, r_{\min}\} : f(x + rw) < f(x)\right\}$;
    if $\bar{r} > -\infty$ **then**
      $x \leftarrow x + \bar{r}w$;
    end
  end
  $T(x) = \mathrm{prox}_{tg}\left(x - t\nabla f(x)\right)$;
  $v \leftarrow \frac{1}{t}(x - T(x)) - \nabla f(x) + \nabla f(T(x))$;
  $x \leftarrow T(x)$;
end

A simple algorithm is described above. Since the correct active manifold $\mathcal{M}$ at the optimal solution is usually unknown, the check "**if** $u \in \mathcal{M}$" is replaced with a suitable problem dependent heuristic. To guard against taking large steps away from the minimizer, a damped linesearch that ensures function value decrease is employed.

Example: Regularized minimization

The ℓ_1-norm

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad (x \in \mathbf{R}^n),$$

is often used as a convex surrogate for sparsity. For example, we may be interested in minimizing a smooth function f, but also require a sparse solution. One way to accomplish this is via the sum-composite optimization problem

$$\min_{x \in \mathbf{R}^n} f(x) + \lambda \|x\|_1$$

for some choice of parameter $\lambda > 0$.

There already exist many approaches to second-order methods for this problem, and the immediate aim is not a new competitive implementation, but rather a simple illustration. Proximal Newton methods [71, 121, 21, 132] incorporate a second-order approximation to the smooth function f in the proximal gradient step. Active-set orthant methods [1, 65, 26] seek to determine the optimal sign of the variables, and minimize quadratic approximations to f over these orthants. The algorithm we derive is most similar to the latter type in being an active-set method, but in contrast to the existing methods our development is much simpler by leveraging the more general framework of this chapter.

Defining the *support* functions

$$\operatorname{supp}(x) = \big\{i \in \{1, \ldots, n\} : x_i \neq 0\big\},$$
$$\operatorname{supp}(x)' = \big\{i \in \{1, \ldots, n\} : x_i = 0\big\},$$

and letting $\{e_i : 1 \leq i \leq n\}$ be the canonical basis of $\mathbf{R}^n$, it is easy to verify that $\|\cdot\|_1$ is partly smooth at $\bar{x}$ relative to the manifold

$$\mathcal{M} = \operatorname{span}\{e_i : i \in \operatorname{supp}(\bar{x})\}.$$

Moreover, locally near $\bar{x}$, $\|\cdot\|_1$ agrees with the smooth function

$$\bar{g}(x) = \sum_{i \in \mathrm{supp}(\bar{x})} \mathrm{sign}(\bar{x}_i) x$$

on the manifold $\mathcal{M}$. Since $\mathcal{M}$ and $\bar{g}$ are defined by linear functions, by Theorem 4.4.1 the second-order subdifferential of $\|\cdot\|_1$ at x for any $v \in \partial \|x\|_1$ is simply

$$\partial^2 \|\cdot\|_1(x|v)(w) = \begin{cases} N_{\mathcal{M}}(x) & w \in T_{\mathcal{M}}(x), \\ \varnothing & w \notin T_{\mathcal{M}}(x). \end{cases}$$

Alternatively,

$$z \in \partial^2 \|\cdot\|_1(x|v)(w) \Leftrightarrow \begin{cases} \mathrm{supp}(w) \subseteq \mathrm{supp}(x), \\ \mathrm{supp}(z) \subseteq \mathrm{supp}(x)', \end{cases}$$

so the Newton step simply amounts to solving the linear system

$$\begin{cases} (\nabla^2 f(x))_i^{\mathsf{T}} w = -v_i & i \in \mathrm{supp}(x), \\ w_i = 0 & i \in \mathrm{supp}(x)'. \end{cases}$$

Stated more simply, we solve

$$\sum_{j \in \mathrm{supp}(x)} \frac{\partial^2 f}{\partial x_i \partial x_j} w_i = -v_i \quad \text{for all } i \in \mathrm{supp}(x), \tag{4.6.5}$$

and set $w_i = 0$ for $i \notin \mathrm{supp}(x)$.

The transversality condition (4.6.2) becomes

$$\begin{cases} \mathrm{supp}(v) \subseteq \mathrm{supp}(\bar{x}), \\ \mathrm{supp}(\nabla^2 f(\bar{x}) v) \subseteq \mathrm{supp}(\bar{x})', \end{cases} \Rightarrow \; v = 0,$$

which holds whenever the matrix of second derivatives

$$\left[\frac{\partial^2 f}{\partial x_i \partial x_j} : (i, j) \in \mathrm{supp}(\bar{x}) \right]$$

is nonsingular at $\bar{x}$.

As an example, we consider the regression model of recovering an s-sparse vector $\hat{x} \in \mathbf{R}^n$ from linear measurements

$$b = A\hat{x} + \epsilon, \quad A \in \mathbf{R}^{m \times n}, \quad b \in \mathbf{R}^m, \quad \epsilon \sim N(0, \sigma^2)$$

via the optimization problem

$$\min_{x \in \mathbf{R}^n} \ \|Ax - b\|_2^2 + \lambda \|x\|_1. \tag{4.6.6}$$

The proximal gradient algorithm, or *iterative soft thresholding* (ISTA), is a popular method for solving (4.6.6) due to its simple iterations. For simplicity we fix the step size $t = \|A^\mathsf{T} A\|_2^{-1}$ (in practice a backtracking line search can be employed to choose the step size). Then the gradient step

$$x - t\nabla f(x) = x - t(A^\mathsf{T} Ax - A^\mathsf{T} b)$$

can be computed in $O(nm)$ operations (ignoring the cost of pre-computing $A^\mathsf{T} A$ and $A^\mathsf{T} b$), and the proximal "shrinkage" step

$$\text{prox}_{t\lambda \|\cdot\|_1}(x)_i = \begin{cases} x_i - t\lambda & (x_i > t\lambda), \\ 0 & (|x_i| \le t\lambda), \\ x_i + t\lambda & (x_i < -t\lambda), \end{cases}$$

in $O(n)$ operations. Letting $k = |\text{supp}(x)|$, the Newton step (4.6.5) involves solving a $k \times k$ linear system, or $O(k^3)$ operations.

While a comprehensive investigation is left as a topic for future research, preliminary numerical experiments demonstrate the effectiveness of incorporating second-order information. Following the experimental setup of [89], consider a randomly generated vector $\hat{x}$ with s nonzero entries, and matrix A with i.i.d.

zero-mean and unit variance entries with dimensions $n = 16s$ and $m = 6s$. If σ is sufficiently small and λ chosen on the order of ϵ, (4.6.6) will have unique nondegenerate minimizer $\bar{x}$ [127]. Since $\nabla^2 f(x) = A^\intercal A$ is positive definite with high probability, the transversality condition holds. We compare ISTA against the accelerated variant FISTA [4] with and without the second-order Newton steps. With 200 nonzero entries for $\hat{x} \in \mathbf{R}^{3200}$, and $m = 1200$ measurements, ISTA and FISTA were both found to take roughly 20 seconds to find the model solution $\bar{x}$ to an accuracy of 10^{-6}. By incorporating second-order subdifferential information and attempting a Newton step when the simple heuristic condition $k^3 \leq mn$ was satisfied, $\bar{x}$ was found up to numerical error in about 3 seconds.

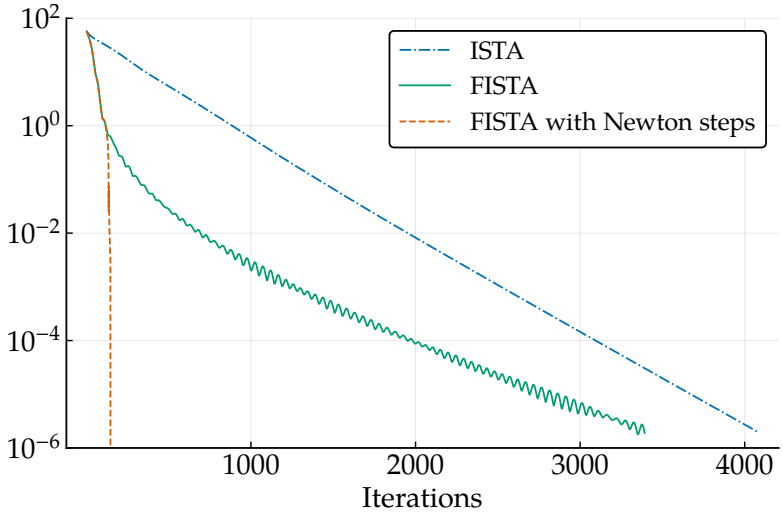


Figure 4.2: $\|x - \bar{x}\|_2$ against iteration count for ISTA and FISTA with and without second-order acceleration on the problem (4.6.6) (accuracy beyond 10^{-6} for the second-order method is not shown).

For a second experiment, consider the regularized logistic regression problem

$$\min_{x \in \mathbf{R}^n} \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y_i d_i^\intercal x}) + \lambda \|x\|_1. \tag{4.6.7}$$

This model is often used to find sparse solutions in binary classification problems, given data vectors $d_1, \ldots, d_m \in \mathbf{R}^n$ and corresponding training labels

$y_1, \ldots, y_m \in \{\pm 1\}$ [105]. We again compare proximal-gradient (ISTA), an accelerated variant (FISTA), and our second-order acceleration scheme on the *colon-cancer* ($n = 2000, m = 62$) dataset from LIBSVM [23], using a standard backtracking line search to choose the step size t. We again observe a fast superlinear rate of convergence for the second-order method once the iterates are sufficiently sparse.

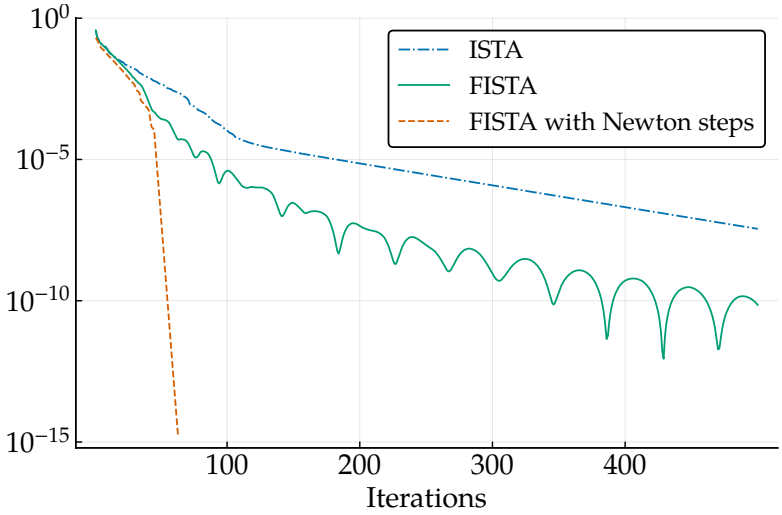


Figure 4.3: $|f(x) - f(\bar{x})|$ against iteration count for ISTA and FISTA with and without second-order acceleration on the problem (4.6.7).

4.7 Composite optimization

One direction for future research is to consider the more general composite optimization problem

$$\min_{x \in \mathbf{R}^n} h(c(x)) \tag{4.7.1}$$

where $h : \mathbf{R}^m \to \bar{\mathbf{R}}$ is closed, convex, proper, and $c : \mathbf{R}^n \to \mathbf{R}^m$ is $C^{(2)}$-smooth. This model has long been recognized as an important class of optimization problems [112, 53, 13, 133, 14, 118, 131, 15, 17, 88], with most of this earlier research driven by examples including nonlinear least-squares, nonlinear programming,

and exact penalty functions. Recently there has been a resurgence of interest in the composite model (4.7.1) due to its importance in a variety of modern applications [83, 41, 46, 45, 43, 25, 24]

Consider a stationary point $0 \in \partial(h \circ c)(\bar{x})$ of the composite objective. By [119, Theorem 10.9] we can apply a chain rule to deduce the existence of a vector $\bar{y}$ satisfying

$$\bar{y} \in \partial h(c(\bar{x})), \qquad \nabla c(\bar{x})^{\mathsf{T}} \bar{y} = 0. \tag{4.7.2}$$

It is well known that h and it's *conjugate*

$$h^*(y) = \sup_{z \in \mathbf{R}^m} \left\{ \langle y, z \rangle - h(z) \right\}$$

satisfy the simple relationship

$$y \in \partial h(c) \Leftrightarrow c \in \partial h^*(y).$$

Therefore we can write the system (4.7.2) as the generalized equation

$$\begin{aligned} 0 &\in \begin{pmatrix} \nabla c(\bar{x})^{\mathsf{T}} \bar{y} \\ -c(\bar{x}) \end{pmatrix} + \begin{pmatrix} \{0\} \\ \partial h^*(\bar{y}) \end{pmatrix} \\ 0 &\in F(\bar{x}, \bar{y}) + \Psi(\bar{x}, \bar{y}) \\ 0 &\in \Phi(\bar{x}, \bar{y}). \end{aligned} \tag{4.7.3}$$

Building on the work of Robinson [114] for the nonlinear programming case, recent manuscripts [28, 16] draw various connections between classical second-order theory, metric regularity, and Newton methods for solving (4.7.3). Our aim is the same in the partly smooth setting.

The map F is smooth with derivative

$$\begin{pmatrix} \sum_{i=1}^m y_i \nabla^2 c_i(x) & \nabla c(x)^{\mathsf{T}} \\ -\nabla c(x) & 0 \end{pmatrix},$$

and when h^* is partly smooth at y for value v, the graphical derivative of Ψ at (x, y) for value $(0, v)$ is

$$D\Psi((x, y)|(0, v)) = \{0\} \times \partial^2 h^*(y|v) = D^*\Psi((x, y)|(0, v)),$$

since Ψ is the subgradient mapping of the partly smooth function $(x, y) \mapsto h^*(y)$.

Now applying the framework of this chapter is straightforward. The transversality condition becomes

$$-\begin{pmatrix} \sum_{i=1}^m \bar{y}_i \nabla^2 c_i(\bar{x}) & \nabla c(\bar{x})^\intercal \\ -\nabla c(\bar{x}) & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \in \begin{pmatrix} 0 \\ \partial^2 h^*(\bar{y}|c(\bar{x}))(v_2) \end{pmatrix} \Rightarrow (v_1, v_2) = (0, 0),$$

and given a point (x, y) and value $F(x, y) + (0, w)$ for $w \in \partial h^*(y)$, the Newton step $d = (d_1, d_2)$ is defined by

$$\begin{pmatrix} -\nabla c(x)^\intercal y \\ c(x) - w \end{pmatrix} \in \begin{pmatrix} \sum_{i=1}^m y_i \nabla^2 c_i(x) & \nabla c(x)^\intercal \\ -\nabla c(x) & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \partial^2 h^*(y|w)(d_2) \end{pmatrix}.$$

An interesting future direction would be to explore settings where we can apply these simple observations. When h is a norm function, it's conjugate is simply the unit ball of the dual norm $|\cdot|_*$, which is partly smooth with simple second-order derivatives in many cases. Are there other interesting settings where ∂h^* is partly smooth? What do the transversality condition and Newton step look like in these scenarios? Do we recover any existing results or algorithms?

CHAPTER 5

A PARTLY SMOOTH NEWTON ALGORITHM

5.1 Introduction

While the theoretical results of the Newton algorithm in Chapter 3 were limited to objective functions with finite max structure, experiments suggest that variants of this method may be effective much more broadly. One particular implementation hurdle arises even for simple nonsmooth functions like the Euclidean norm: solving the quadratic subproblem directly will be numerically unstable, due to the ill-conditioning of the Hessians $\nabla^2 f(s)$ when s is close to zero.

However, for partly smooth functions (Definition 2.5.4) this ill-conditioning is typically highly structured. Informally, when $f : \mathbf{R}^n \to \mathbf{R}$ is partly smooth at $\bar{x}$ relative to active manifold $\mathcal{M}$, the function behaves smoothly along directions tangent to $\mathcal{M}$, and in a nonsmooth manner or "sharply" along directions normal to $\mathcal{M}$. Experimentally results suggest that for s close to $\bar{x}$, the nonsmooth subspace $N_{\mathcal{M}}(\bar{x})$ is approximately spanned by the eigenvectors corresponding to large eigenvalues of the Hessian $\nabla^2 f(s)$. When projected onto $T_{\mathcal{M}}(\bar{x})$, the Hessian is well-conditioned. Motivated by this idea, we follow a simple strategy for solving the system (3.7.1), similar to reduced system approaches for nonlinear programming described in standard texts [106, 9] and avoiding full Hessian computations.

5.2 A reduced system approach

Given function $f : \mathbf{R}^n \to \mathbf{R}$ that is $C^{(2)}$-smooth around every point in $\mathcal{D} \subset \mathbf{R}^n$, recall the local linear and quadratics approximations for $s \in \mathcal{D}$

$$\begin{aligned} l_s(\cdot) &= f(s) + \langle \nabla f(s), \cdot - s \rangle, \\ q_s(\cdot) &= l_s(\cdot) + \frac{1}{2} \langle \cdot - s, \nabla^2 f(s)(\cdot - s) \rangle. \end{aligned}$$

Given a finite bundle $S \subset \mathcal{D}$, our Newton algorithms of Chapter 3 solve the subproblem

$$\begin{aligned} \underset{t \in \mathbf{R},\, x \in \mathbf{R}^n}{\text{minimize}} \quad & \sum_s \lambda_s q_s(x) \\ \text{subject to} \quad & l_s(x) \text{ equal for all } s \in S. \end{aligned}$$

Let G and b be a matrix and vector satisfying

$$\{x \in \mathbf{R}^n : Gx = b\} = \{x \in \mathbf{R}^n : l_s(x) \text{ equal for all } s \in S\}.$$

Assuming the gradients $\nabla f(S)$ are affinely independent, we can in particular find a G that is full rank, and we can write the optimality conditions of the subproblem as

$$\begin{aligned} \sum_{s \in S} \lambda_s \nabla^2 f(s)(x - s) + G^T \nu &= -\sum_{s \in S} \lambda_s \nabla f(s) \\ Gx &= b, \end{aligned} \tag{5.2.1}$$

for multiplier vector $\nu \in \mathbf{R}^{|S|-1}$. Suppose that we have also found matrices U and V such that the matrix $\begin{bmatrix} U & V \end{bmatrix} \in \mathbf{R}^{n \times n}$ is full rank and $GU = 0$ (via a QR factorization of G^T, for example). The columns of U are then a basis for Null(G), and we can write any solution of (5.2.1) as

$$x = Ux_u + Vx_v.$$

The constraint $Gx = b$ then implies $GVx_v = b$, which can be solved for x_v, since G (and hence GV) is full rank. We deduce

$$\{x : Gx = b\} = \text{Range}(U) + p,$$

where p is the particular solution $V(GV)^{-1}b$. Substituting this into the stationarity condition and multiplying through by U^T yields the *reduced* system

$$\sum_{s\in S} \lambda_s U^\mathsf{T}\nabla^2 f(s)(Ux_u + p - s) = -\sum_{s\in S} \lambda_s U^\mathsf{T}\nabla f(s).$$

In a slight modification to the algorithm, if we project each reference point onto the active subspace we arrive at the linear system

$$\sum_{s\in S} \lambda_s U^\mathsf{T}\nabla^2 f(s)Ux_u = \sum_{s\in S} \lambda_s \left[(U^\mathsf{T}\nabla^2 f(s)U)U^\mathsf{T}(s-p) - U^\mathsf{T}\nabla f(s)\right]. \tag{5.2.2}$$

This system only involves the *projected Hessians* $U^\mathsf{T}\nabla^2 f(s)U$, which remain well-conditioned if the span of V is close to the subspace $N_{\mathcal{M}}(\bar{x})$, a property experimentally observed to hold in practice.

A simple test involves "mixed norm" functions of the form

$$f(x) = \sqrt{x^\mathsf{T}Ax} + x^\mathsf{T}Bx \qquad (x \in \mathbf{R}^n) \tag{5.2.3}$$

for positive definite matrices $A, B \in \mathbf{R}^{n\times n}$. First introduced to study BFGS [81], the function was recently reexamined in the context of $\mathcal{VU}$-theory in the ICM lecture [120]. The function is partly smooth with respect to the manifold $\mathcal{M} = \text{Null}(A)$, and twice continuously differentiable on the open set $\mathbf{R}^n \setminus \mathcal{M}$, with a unique minimizer of 0. Despite the impossibility of writing (5.2.3) as a maximum of finite smooth functions, the reduced system modification to Algorithm 3.1 is highly effective, illustrated in Figure 5.1. We observe that after using the iterates of a first-order bundle method to estimate $\dim \mathcal{M}$, superlinear convergence to the minimizer is possible by incorporating the reduced second-order information.

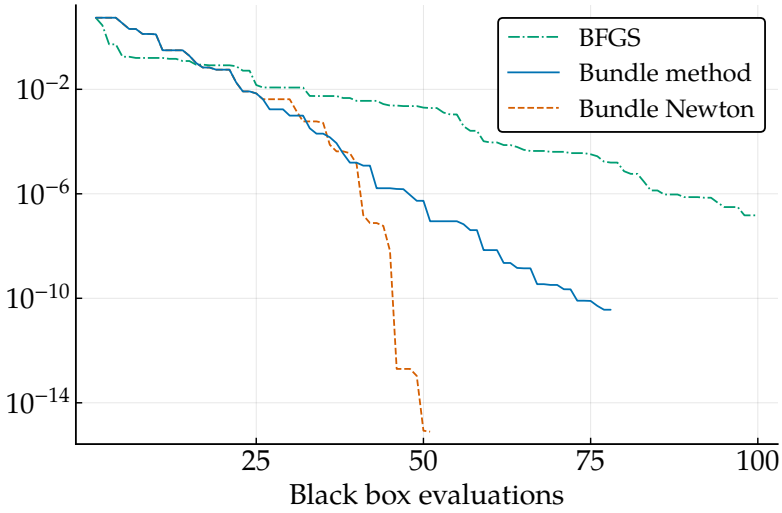


Figure 5.1: Best function value found for BFGS, a bundle method, and the reduced bundle Newton algorithm against number of black box evaluations on the function (5.2.3) with $A = \operatorname{diag}(1, 0, 1, 0, \ldots)$ and $B = (1, 1/2^2, \ldots, 1/n^2)$ in dimension $n = 8$.

5.3 Example: Eigenvalue optimization

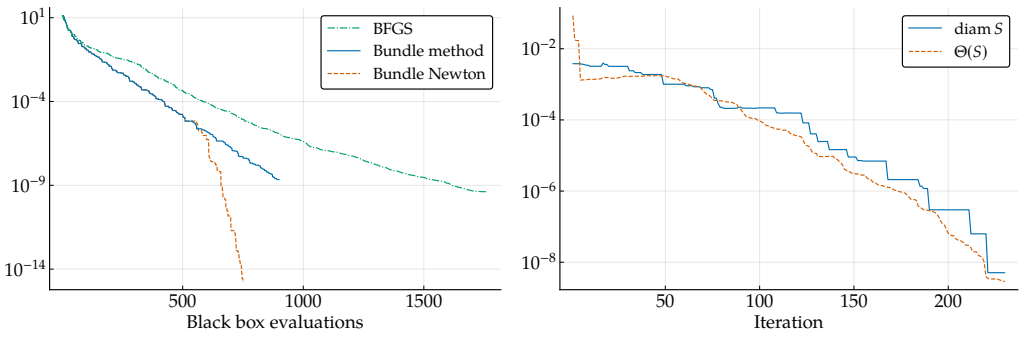
Our final experiment is an eigenvalue problem. Specifically, given symmetric matrices $A_0, \ldots, A_n \in \mathbf{R}^{m \times m}$ we seek to minimize

$$f(x) = \lambda_{\max}\Big(A_0 + \sum_{i=1}^{n} x_i A_i\Big), \qquad (x \in \mathbf{R}^n) \tag{5.3.1}$$

where $\lambda_{\max}(\,\cdot\,)$ is the largest eigenvalue function. Typically minimizers occur at points where $\lambda_{\max}$ has multiplicity $t > 1$, necessitating nonsmooth minimization techniques. Under reasonable conditions, the set of points for which $\lambda_{\max}$ has fixed multiplicity t is a manifold of codimension $\frac{t(t+1)}{2}$, relative to which f is partly smooth (see e.g., [79]).

For illustration, Figure 5.2 shows convergence of the bundle method, BFGS, and our Newton method on a typical trial for this problem using random data. All algorithms were run without termination conditions until numerical issues prevented any further progress. In this example for $n = 50$ matrices in $\mathbf{R}^{25\times 25}$, the

optimal eigenvalue multiplicity was 6, and we again observe fast convergence of the bundle Newton algorithm once the subdifferential dimension $\frac{t(t+1)}{2} - 1 = 20$ can be identified.



(a) $f(x) - \min f$ (b) diam S and $\Theta(S)$

Figure 5.2: Function value convergence and optimality measures for the maximum eigenvalue function (5.3.1) for $n = 50$ random symmetric matrices in $\mathbf{R}^{25\times 25}$.

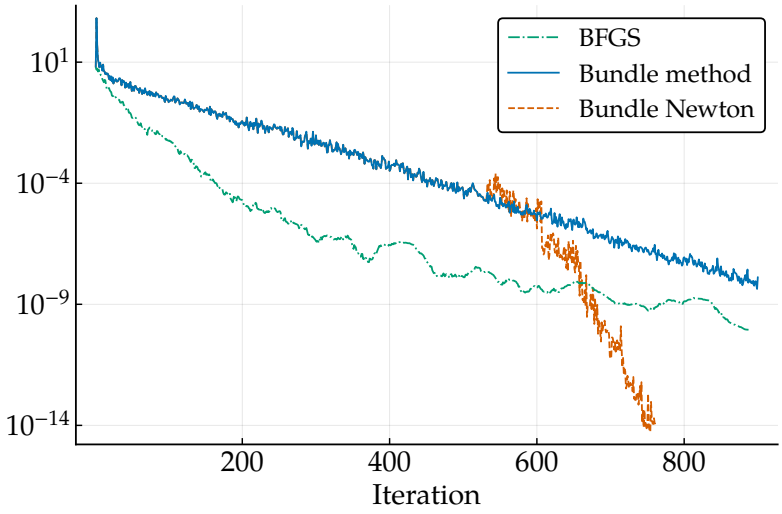


Figure 5.3: Clustering of the six largest eigenvalues for the random trial depicted in Figure 5.2.

(Note that since the optimal objective value is unknown, the figures were generated using the best objective value found after running the algorithms with a large number of random starting points. This introduces a slight bias in the accuracy reported for the bundle Newton algorithm.) In Figure 5.3, we observe that

the bundle Newton algorithm achieves an eigenvalue clustering several orders of magnitude better than is possible with a bundle method or BFGS.

Using the active manifold to accelerate eigenvalue optimization is not new [109, 123, 108]. What is remarkable is that the bundle Newton algorithm, combined with a first phase such as a traditional bundle method, rapidly convergences to the minimizer *without* any structural knowledge of the function.

5.4 Future directions

Development of a complete convergence theory in the partly smooth setting is a topic for future research. For max functions

$$f(x) = \max_{i=1,\ldots,k} f_i(x) \qquad (x \in \mathbf{E})$$

with smooth components, the strong second-order conditions (3.5.1) for $\bar{x}$ are exactly that $0 \in \operatorname{ri} \partial f(\bar{x})$ with ∂f partly smooth at $\bar{x}$ for 0 relative to the manifold

$$\mathcal{M} = \{x \in \mathbf{E} : f_i(x) \text{ equal for all } i\}.$$

An immediate question is what is the correct generalization of a full bundle from max functions to partly smooth functions? A convergence proof for the Newton algorithm described above would likely need to combine ideas from Chapter 4 with the proof techniques of Chapter 3.

Also, our observation about the connections between partial smoothness and structured ill-conditioning of Hessians is only anecdotal. A more rigorous theory would be interesting. To what extent can Hessians of functions that are $C^{(2)}$-smooth almost everywhere illuminate partly smooth structure?

BIBLIOGRAPHY

[1] G. Andrew and J. Gao, Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ACM, 2007, 33–40.

[2] D. S. Atkinson and P. M. Vaidya, A cutting plane algorithm for convex programming that uses analytic centers. *Mathematical Programming* **69** (1995) 1–43.

[3] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011.

[4] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** (2009) 183–202.

[5] D. P. Bertsekas, *Convex Optimization Algorithms*. Athena Scientific, Belmont, MA, 2015.

[6] P. T. Boggs and J. W. Tolle, Sequential quadratic programming. *Acta Numerica* **4** (1995) 1–51.

[7] J. Bolte, A. Daniilidis, and A. Lewis, Tame functions are semismooth. *Mathematical Programming* **117** (2009) 5–19.

[8] J. F. Bonnans, Local analysis of Newton-type methods for variational inequalities and nonlinear programming. *Applied Mathematics and Optimization* **29** (1994) 161–186.

[9] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*. Springer-Verlag, Berlin, 2006.

[10] L. Bottou, F. E. Curtis, and J. Nocedal, Optimization methods for large-scale machine learning. *SIAM Review* **60** (2018) 223–311.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** (2011) 1–122.

[12] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. A. Simões, Gradient sampling methods for nonsmooth optimization. In *Numerical Nonsmooth Optimization*, Springer, to appear, 2019.

[13] J. V. Burke, Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming* **33** (1985) 260–279.

[14] J. V. Burke, Second order necessary and sufficient conditions for convex composite NDO. *Mathematical Programming* **38** (1987) 287–302.

[15] J. V. Burke and R. Poliquin, Optimality conditions for non-finite valued convex composite functions. *Mathematical Programming* **57** (1992) 103–120.

[16] J. V. Burke and A. Engle, Strong metric (sub) regularity of KKT mappings for piecewise linear-quadratic convex-composite optimization. *arXiv:1805.01073 [math.OC]* (2018).

[17] J. V. Burke and M. C. Ferris, A Gauss-Newton method for convex composite optimization. *Mathematical Programming* **71** (1995) 179–194.

[18] J. V. Burke, A. S. Lewis, and M. L. Overton, A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization* **15** (2005) 751–779.

[19] J. V. Burke and J. J. Moré, On the identification of active constraints. *SIAM Journal on Numerical Analysis* **25** (1988) 1197–1211.

[20] J. Burke, On the identification of active constraints II: The nonconvex case. *SIAM Journal on Numerical Analysis* **27** (1990) 1081–1102.

[21] R. H. Byrd, J. Nocedal, and F. Öztoprak, An inexact successive quadratic approximation method for L-1 regularized optimization. *Mathematical Programming* **157** (2016) 375–396.

[22] P. H. Calamai and J. J. Moré, Projected gradient methods for linearly constrained problems. *Mathematical Programming* **39** (1987) 93–116.

[23] C. C. Chang and C. J. Lin, LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2** (2011) 1–27.

[24] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv:1904.10020 [math.OC]* (2019).

[25] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy, Composite optimization for robust blind deconvolution. *arXiv:1901.01624 [math.OC]* (2019).

[26] T. Chen, F. E. Curtis, and D. P. Robinson, FaRSA for ℓ1-regularized convex optimization: local convergence and numerical experience. *Optimization Methods and Software* **33** (2018) 396–415.

[27] E. W. Cheney and A. A. Goldstein, Newton's method for convex programming and Tchebycheff approximation. *Numerische Mathematik* **1** (1959) 253–268.

[28] R. Cibulka, A. Dontchev, and A. Kruger, Strong metric subregularity of mappings in variational analysis and optimization. *Journal of Mathematical Analysis and Applications* **457** (2018) 1247–1282.

[29] F. H. Clarke, *Optimization and Nonsmooth Analysis*. Wiley, New York, 1983.

[30] P. L. Combettes and J. C. Pesquet, Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, Springer, 2011, 185–212.

[31] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-free Optimization*. SIAM, Philadelphia, 2009.

[32] F. E. Curtis, T. Mitchell, and M. L. Overton, A BFGS-SQP method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles. *Optimization Methods and Software* **32** (2017) 148–181.

[33] F. E. Curtis and X. Que, An adaptive gradient sampling algorithm for non-smooth optimization. *Optimization Methods and Software* **28** (2013) 1302–1324.

[34] F. E. Curtis, D. P. Robinson, and B. Zhou, A self-correcting variable-metric algorithm framework for nonsmooth optimization. *IMA Journal of Numerical Analysis,* to appear (2019).

[35] A. Daniilidis, C. Sagastizábal, and M. Solodov, Identifying structure of nonsmooth convex functions by the bundle technique. *SIAM Journal on Optimization* **20** (2009) 820–840.

[36] D. Davis and D. Drusvyatskiy, Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization* **29** (2019) 207–239.

[37] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications* **179** (2018) 962–982.

[38] A. L. Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings*. Springer-Verlag, New York, 2014.

[39] A. L. Dontchev and R. T. Rockafellar, Regularity and conditioning of solution mappings in variational analysis. *Set-Valued Analysis* **12** (2004) 79–109.

[40] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization* **26** (2016) 513–534.

[41] D. Drusvyatskiy and A. S. Lewis, Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research* **43** (2018) 919–948.

[42] D. Drusvyatskiy and A. S. Lewis, Optimality, identifiability, and sensitivity. *Mathematical Programming* **147** (2014) 467–498.

[43] D. Drusvyatskiy and C. Paquette, Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming* **178** (2019) 503–558.

[44] Y. Du and A. Ruszczyński, Rate of convergence of the bundle method. *Journal of Optimization Theory and Applications* **173** (2017) 908–922.

[45] J. C. Duchi and F. Ruan, Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference* **8** (2018) 471–529.

[46] J. C. Duchi and F. Ruan, Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization* **28** (2018) 3229–3259.

[47] J. C. Dunn, On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications* **55** (1987) 203–216.

[48] F. Facchinei, A. Fischer, and C. Kanzow, On the accurate identification of active constraints. *SIAM Journal on Optimization* **9** (1998) 14–32.

[49] F. Facchinei and J. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag, New York, 2003.

[50] M. C. Ferris, Finite termination of the proximal point algorithm. *Mathematical Programming* **50** (1991) 359–366.

[51] A. Fischer, Local behavior of an iterative framework for generalized equations with nonisolated solutions. *Mathematical Programming* **94** (2002) 91–124.

[52] S. D. Flåm, On finite convergence and constraint identification of subgradient projection methods. *Mathematical Programming* **57** (1992) 427–437.

[53] R. Fletcher, Second order corrections for non-differentiable optimization. In *Numerical Analysis*, Springer, 1982, 85–114.

[54] A. Fuduli, M. Gaudioso, and G. Giallombardo, Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization* **14** (2004) 743–756.

[55] H. Gfrerer and J. V. Outrata, On a semismooth* Newton method for solving generalized equations. *arXiv:1904.09167 [math.OC]* (2019).

[56] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2013.

[57] W. Hare and C. Sagastizábal, A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization* **20** (2010) 2442–2473.

[58] W. Hare and C. Sagastizábal, Computing proximal points of nonconvex functions. *Mathematical Programming* **116** (2009) 221–258.

[59] E. S. Helou, S. A. Santos, and L. E. Simões, A fast gradient and function sampling method for finite-max functions. *Computational Optimization and Applications* **71** (2018) 673–717.

[60] E. S. Helou, S. A. Santos, and L. E. Simões, On the local convergence analysis of the gradient sampling method for finite max-functions. *Journal of Optimization Theory and Applications* **175** (2017) 137–157.

[61] A. F. Izmailov and M. V. Solodov, *Newton-type Methods for Optimization and Variational Problems*. Springer, Cham, 2014.

[62] A. F. Izmailov and M. V. Solodov, Newton-type methods: a broader view. *Journal of Optimization Theory and Applications* **164** (2015) 577–620.

[63] N. H. Josephy, *Newton's method for generalized equations*. Tech. rep. Mathematics Research Center, University of Wisconsin, Madison, 1979.

[64] J. E. Kelley Jr, The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics* **8** (1960) 703–712.

[65] N. Keskar, J. Nocedal, F. Öztoprak, and A. Waechter, A second-order method for convex ℓ_1-regularized optimization with active-set prediction. *Optimization Methods and Software* **31** (2016) 605–621.

[66] F. Al-Khayyal and J. Kyparisis, Finite convergence of algorithms for nonlinear programs and variational inequalities. *Journal of Optimization Theory and Applications* **70** (1991) 319–332.

[67] K. C. Kiwiel, Efficiency of proximal bundle methods. *Journal of Optimization Theory and Applications* **104** (2000) 589–603.

[68] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*. Springer-Verlag, Berlin, 1985.

[69] K. C. Kiwiel, A linearization algorithm for nonsmooth minimization. *Mathematics of Operations Research* **10** (1985) 185–194.

[70] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization*. Kluwer, Dordrecht, 2002.

[71] J. D. Lee, Y. Sun, and M. A. Saunders, Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization* **24** (2014) 1420–1443.

[72] J. M. Lee, *Introduction to Smooth Manifolds*. Springer, New York, 2003.

[73] S. Lee and S. J. Wright, Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research* **13** (2012) 1705–1744.

[74] Y. T. Lee, A. Sidford, and S. C. Wong, A faster cutting plane method and its implications for combinatorial and convex optimization. In *56th Annual Symposium on Foundations of Computer Science*, IEEE, 2015, 1049–1065.

[75] C. Lemaréchal, An extension of Davidon methods to non differentiable problems. In *Nondifferentiable optimization*, Springer, 1975, 95–109.

[76] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, New variants of bundle methods. *Mathematical Programming* **69** (1995) 111–147.

[77] C. Lemaréchal, F. Oustry, and C. Sagastizábal, The $\mathcal{U}$-Lagrangian of a convex function. *Transactions of the American Mathematical Society* **352** (2000) 711–729.

[78] C. Lemaréchal and C. Sagastizábal, Variable metric bundle methods: From conceptual to implementable forms. *Mathematical Programming* **76** (1997) 393–410.

[79] A. S. Lewis, Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization* **13** (2002) 702–725.

[80] A. S. Lewis and J. Liang, Partial smoothness and constant rank. *arXiv:1807.03134 [math.OC]* (2018).

[81] A. S. Lewis and M. L. Overton, Nonsmooth optimization via BFGS. http://cs.nyu.edu/overton/papers/pdffiles/bfgs_inexactLS.pdf (2008).

[82] A. S. Lewis and M. L. Overton, Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming* **141** (2013) 135–163.

[83] A. S. Lewis and S. J. Wright, A proximal method for composite minimization. *Mathematical Programming* **158** (2016) 501–546.

[84] A. S. Lewis and S. J. Wright, Identifying activity. *SIAM Journal on Optimization* **21** (2011) 597–614.

[85] A. S. Lewis and C. J. S. Wylie, Active-set Newton methods and partial smoothness. *arXiv:1902.00724 [math.OC]* (2019).

[86] A. S. Lewis and S. Zhang, Partial smoothness, tilt stability, and generalized Hessians. *SIAM Journal on Optimization* **23** (2013) 74–94.

[87] A. Lewis and C. J. S. Wylie, A simple Newton method for local nonsmooth optimization. *arXiv:1907.11742 [math.OC]* (2019).

[88] C. Li and X. Wang, On convergence of the Gauss-Newton method for convex composite optimization. *Mathematical Programming* **91** (2002) 349–356.

[89] J. Liang, J. Fadili, and G. Peyré, Activity Identification and Local Linear Convergence of Forward–Backward-type Methods. *SIAM Journal on Optimization* **27** (2017) 408–437.

[90] J. Liang, J. Fadili, and G. Peyré, Local linear convergence analysis of Primal–Dual splitting methods. *Optimization* **67** (2018) 821–853.

[91] L. Lukšan and J. Vlček, A bundle-Newton method for nonsmooth unconstrained minimization. *Mathematical Programming* **83** (1998) 373–391.

[92] R. Mifflin, A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization. In *Nondifferential and Variational Techniques in Optimization*, Springer, 1982, 77–90.

[93] R. Mifflin, An algorithm for constrained optimization with semismooth functions. *Mathematics of Operations Research* **2** (1977) 191–207.

[94] R. Mifflin and C. Sagastizábal, $\mathcal{VU}$-smoothness and proximal point results for some nonconvex functions. *Optimization Methods and Software* **19** (2004) 463–478.

[95] R. Mifflin and C. Sagastizábal, A $\mathcal{VU}$-algorithm for convex minimization. *Mathematical Programming* **104** (2005) 583–608.

[96] R. Mifflin and C. Sagastizábal, A science fiction story in nonsmooth optimization originating at IIASA. *Documenta Mathematica* Extra Volume: Optimization Stories (2012) 291–300.

[97] R. Mifflin and C. Sagastizábal, On $\mathcal{VU}$-theory for functions with primal-dual gradient structure. *SIAM Journal on Optimization* **11** (2000) 547–571.

[98] S. A. Miller and J. Malick, Newton methods for nonsmooth convex minimization: connections among $\mathcal{U}$-Lagrangian, Riemannian Newton and SQP methods. *Mathematical Programming* **104** (2005) 609–633.

[99] B. S. Mordukhovich, Sensitivity analysis in nonsmooth optimization. In *Theoretical Aspects of Industrial Design*, SIAM Philadelphia, 1992, 32–46.

[100] J. J. Moreau, Fonctions convexes duales et points proximaux dans un espace hilbertien. *C. R. Acad. Sci. Paris Sér. A Math.* **255** (1962) 2897–2899.

[101] J. J. Moreau, Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France* **93** (1965) 273–299.

[102] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

[103] Y. Nesterov, Complexity estimates of some cutting plane methods based on the analytic barrier. *Mathematical Programming* **69** (1995) 149–176.

[104] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Kluwer, Boston, 2004.

[105] A. Y. Ng, Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the 21st International conference on Machine Learning*, ACM, 2004, 78–85.

[106] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, New York, 2006.

[107] W. de Oliveira and C. Sagastizábal, Bundle methods in the XXIst century: A bird's-eye view. *Pesquisa Operacional* **34** (2014) 647–670.

[108] F. Oustry, A second-order bundle method to minimize the maximum eigenvalue function. *Mathematical Programming* **89** (2000) 1–33.

[109] M. L. Overton, On minimizing the maximum eigenvalue of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications* **9** (1988) 256–268.

[110] M. L. Overton and R. S. Womersley, Second derivatives for optimizing eigenvalues of symmetric matrices. *SIAM Journal on Matrix Analysis and Applications* **16** (1995) 697–718.

[111] N. Parikh and S. Boyd, Proximal algorithms. *Foundations and Trends in Optimization* **1** (2014) 127–239.

[112] M. J. Powell, Algorithms for nonlinear constraints that use Lagrangian functions. *Mathematical Programming* **14** (1978) 224–248.

[113] L. Qi and J. Sun, A nonsmooth version of Newton's method. *Mathematical Programming* **58** (1993) 353–367.

[114] S. M. Robinson, Strongly regular generalized equations. *Mathematics of Operations Research* **5** (1980) 43–62.

[115] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[116] R. T. Rockafellar, Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* **14** (1976) 877–898.

[117] R. T. Rockafellar, Proto-differentiability of set-valued mappings and its applications in optimization. *Annales de l'Institut Henri Poincaré C, Analyse Non Linéaire* **6** (1989) 449–482.

[118] R. T. Rockafellar, Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives. *Mathematics of Operations Research* **14** (1989) 462–484.

[119] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Springer-Verlag, Berlin, 2009.

[120] C. Sagastizábal, A $\mathcal{VU}$-point of view of nonsmooth optimization. In *Proceedings of the International Congress of Mathematicians, Rio de Janeiro*, vol. 3. 2018, 3785–3806.

[121] K. Scheinberg and X. Tang, Practical inexact proximal quasi-Newton method with global complexity analysis. *Mathematical Programming* **160** (2016) 495–529.

[122] H. Schramm and J. Zowe, A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization* **2** (1992) 121–152.

[123] A. Shapiro and M. K. Fan, On eigenvalue optimization. *SIAM Journal on Optimization* **5** (1995) 552–569.

[124] N. Z. Shor, Utilization of the operation of space dilatation in the minimization of convex functions. *Cybernetics and Systems Analysis* **6** (1972) 7–15.

[125] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. SIAM, Philadelphia, 2011.

[126] P. M. Vaidya, A new algorithm for minimizing convex functions over convex sets. *Mathematical Programming* **73** (1996) 291–341.

[127] S. Vaiter, G. Peyré, and J. Fadili, Model consistency of partly smooth regularizers. *IEEE Transactions on Information Theory* **64** (2017) 1725–1737.

[128] J. Vlček and L. Lukšan, Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. *Journal of Optimization Theory and Applications* **111** (2001) 407–430.

[129] P. Wolfe, A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable Optimization*, Springer, 1975, 145–173.

[130] S. J. Wright, Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization* **31** (1993) 1063–1079.

[131] S. J. Wright, Local properties of inexact methods for minimizing nonsmooth composite functions. *Mathematical Programming* **37** (1987) 232.

[132] G. X. Yuan, C. H. Ho, and C. J. Lin, An improved GLMNET for L1-regularized logistic regression. *Journal of Machine Learning Research* **13** (2012) 1999–2030.

[133] Y. Yuan, On the superlinear convergence of a trust region algorithm for nonsmooth optimization. *Mathematical Programming* **31** (1985) 269–285.