Math. Program., Ser. B 104, 609–633 (2005)

Digital Object Identifier (DOI) 10.1007/s10107-005-0631-2

Scott A. Miller · Jérôme Malick

Newton methods for nonsmooth convex minimization: connections among $\mathcal{U}$-Lagrangian, Riemannian Newton and SQP methods

This paper is dedicated to R.T. Rockafellar, on the occasion of his 70th birthday.

Received: May 5, 2004 / Accepted: March 15, 2005
Published online: July 14, 2005 – © Springer-Verlag 2005

Abstract. This paper studies Newton-type methods for minimization of partly smooth convex functions. Sequential Newton methods are provided using local parameterizations obtained from $\mathcal{U}$-Lagrangian theory and from Riemannian geometry. The Hessian based on the $\mathcal{U}$-Lagrangian depends on the selection of a dual parameter g; by revealing the connection to Riemannian geometry, a natural choice of g emerges for which the two Newton directions coincide. This choice of g is also shown to be related to the least-squares multiplier estimate from a sequential quadratic programming (SQP) approach, and with this multiplier, SQP gives the same search direction as the Newton methods.

1. Introduction

1.1. Motivation

Newton's method is the canonical fast optimization algorithm. For a smooth function f, Newton's method converges quadratically to a stationary point $\bar{x}$ of f when $\nabla^2 f(\bar{x})$ is nonsingular. Furthermore, the spirit of Dennis and Moré's celebrated quasi-Newton result [4] is that, for a strictly convex smooth function, a convergent variable-metric steepest descent algorithm converges superlinearly to the minimum *if and only if* the algorithm appears "Newton-like" along search directions in the limit. Given the importance of Newton's method for fast convergence, one would like to extend it to nonsmooth functions.

This paper compares several approaches to defining a Newton method for minimizing a nonsmooth convex function $f : \mathbb{R}^n \to \mathbb{R}$ with an additional structure: f is assumed to be partly smooth in the sense of Lewis [12]. Roughly speaking, this means that there exists a smooth manifold $\mathcal{M}$ in which f is smooth, and normal to which f is not differentiable. As explained in [12] (see also [23]), two important classes of partly smooth functions are finite max-functions and maximum eigenvalue functions. Section 2 studies the particular properties of the $\mathcal{U}$-Lagrangian theory [11] under the partial smoothness assumption.

S. A. Miller: Numerica Corp., P.O. Box 271246, Ft. Collins, CO 80527-1246, USA.
e-mail: samiller@numerica.us

J. Malick: INRIA, 655 avenue de l'Europe, 38334 Saint Ismier Cedex, France.
e-mail: jerome.malick@inria.fr

If we also assume that a minimizer of f belongs to $\mathcal{M}$, minimizing f on $\mathbb{R}^n$ amounts to minimizing f on $\mathcal{M}$:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathcal{M}. \end{array} \tag{1.1}$$

This seems to complicate the situation. On the contrary: restricted to $\mathcal{M}$ everything is smooth, and methods of smooth optimization can then be adapted. But what does it mean to apply Newton's method on a manifold? Since a p-dimensional manifold locally resembles $\mathbb{R}^p$, we can consider a smooth local parameterization $\varphi : \mathbb{R}^p \to \mathcal{M}$ and apply the usual Newton iteration to the composition $f \circ \varphi$. One choice of parameterization derived from the $\mathcal{U}$-Lagrangian leads to the $\mathcal{U}$-Newton method studied in Section 3. Another choice derived from Riemannian geometry leads to the Riemannian Newton method studied in [26, 6, 3] in particular. Section 4 presents the connections between these two methods: with the right selection of the dual parameter g for $\mathcal{U}$-Newton, both methods give the same Newton direction and converge quadratically.

A second interpretation of (1.1) gives rise to a sequential quadratic programming (SQP) approach. Replace f with a smooth function $\bar{f}$ that agrees with f on $\mathcal{M}$, and describe $\mathcal{M}$ by smooth equations $\Phi : \mathbb{R}^n \to \mathbb{R}^{n-p}$, so that (1.1) is locally equivalent to the smooth constrained problem

$$\begin{array}{ll} \min & \bar{f}(x) \\ \text{s.t.} & \Phi(x) = 0. \end{array} \tag{1.2}$$

SQP uses Newton's method to solve the optimality conditions of (1.2), generating a quadratic program to solve at each step. This is the idea behind the second-order methods for eigenvalue optimization in [22, 25]. Section 5 presents the connections between $\mathcal{U}$-Newton and SQP methods. Just as the $\mathcal{U}$-Newton method depends on a choice of $g \in \partial f(x)$ which defines the $\mathcal{U}$-Lagrangian, SQP depends on the choice of approximate Lagrange multipliers λ. We will show that the choice of g leading to quadratic convergence of $\mathcal{U}$-Newton corresponds to the selection of least-squares multipliers for λ. Moreover, this choice makes the Newton and SQP directions identical.

1.2. Notation and assumptions

Basic notation. For a subset $S \subset \mathbb{R}^n$, lin S, aff S and ri S denote the linear hull, affine hull and relative interior of S, respectively. When it is well-defined, $\mathrm{P}_S(x)$ is the projection of x onto S. The closed ball $\{y \in \mathbb{R}^n \mid \|y - x\| \leq r\}$ is written $B(x, r)$.

Differential geometry. Roughly speaking, a sub-manifold in $\mathbb{R}^n$ is a set consisting locally of the solutions of some smooth equations with linearly independent gradients. Precisely, a subset $\mathcal{M}$ of $\mathbb{R}^n$ is said to be *a p-dimensional differentiable sub-manifold* of class C^k around $x \in \mathcal{M}$ ($k \in \mathbb{N} \cup \{\infty\}$) if there is a C^k-function $\Phi : \mathbb{R}^n \to \mathbb{R}^{n-p}$ such that, for all y close enough to x,

$$y \in \mathcal{M} \iff \Phi(y) = 0,$$

and in addition the derivative of Φ at x is surjective. We say that $\Phi(y) = 0$ is a *local equation* of $\mathcal{M}$ near x. For a sub-manifold $\mathcal{M}$, we denote respectively by $\mathrm{T}_{\mathcal{M}}(x)$ and $\mathrm{N}_{\mathcal{M}}(x)$ the tangent and normal subspaces to $\mathcal{M}$ at $x \in \mathcal{M}$. The tangent bundle $\mathrm{T}\mathcal{M}$ and normal bundle $\mathrm{N}\mathcal{M}$ are defined by

$$\mathrm{T}\mathcal{M} = \bigcup_{x \in \mathcal{M}} (x, \mathrm{T}_{\mathcal{M}}(x)), \qquad \mathrm{N}\mathcal{M} = \bigcup_{x \in \mathcal{M}} (x, \mathrm{N}_{\mathcal{M}}(x)).$$

These are C^{k-1}-sub-manifolds of $\mathbb{R}^{2n}$ of dimension $2p$ and n, respectively.

When F is a differentiable function between two C^1-sub-manifolds X and Y, the derivative $\mathrm{D}F(x)$ is a linear map from $\mathrm{T}_X(x)$ to $\mathrm{T}_Y(F(x))$.

Definition 1.1. *Let $\mathcal{M}$ be a p-dimensional C^∞-manifold and $x \in \mathcal{M}$. A function φ_x is said to be a* local parameterization of $\mathcal{M}$ around x *if there exist a neighborhood Θ of 0 in $\mathrm{T}_{\mathcal{M}}(x)$ and a neighborhood Ω of x in $\mathcal{M}$, such that $\varphi_x : \Theta \to \Omega$ is a C^∞-diffeomorphism ($\varphi_x : \Theta \to \Omega$ is a bijection, and φ_x and φ_x^{-1} are of class C^∞), and $\varphi_x(0) = x$.*

Definition 1.2. *We say that a family $\{\varphi_x\}_x$ is a* smooth parameterization family of $\mathcal{M}$ *if φ_x is a local parameterization around x and if the function $(x, \eta) \mapsto \varphi_x(\eta)$ from $\mathrm{T}\mathcal{M}$ to $\mathcal{M}$ is C^∞.*

The three parameterization families that we consider in this paper (tangential parameterization, exponential parameterization and projection parameterization) are smooth (see respectively Lemmas 3.3, 4.1 and 4.8).

Partial smoothness. Lewis introduced the notion of partly smooth functions in [12]. This concept expresses an underlying smooth structure of a nonsmooth function.

Definition 1.3. *A function f is a C^k-*partly smooth *at x relative to $\mathcal{M} \subset \mathbb{R}^n$ ($k \in \mathbb{N} \cup \{\infty\}$), if $\mathcal{M}$ is a C^k-sub-manifold around $x \in \mathcal{M}$ and the following properties hold:*

(i) restricted smoothness*: the restriction of f to $\mathcal{M}$ is a C^k-function near x;*
(ii) regularity*: f is Clarke regular [24] at all $y \in \mathcal{M}$ near x, with $\partial f(y) \neq \emptyset$;*
(iii) normal sharpness*: for any $d \in \mathrm{N}_{\mathcal{M}}(x)$, the function $t \mapsto f(x + td)$ is not differentiable at $t = 0$;*
(iv) subdifferential continuity*: the set-valued map ∂f restricted to $\mathcal{M}$ is continuous at x.*

Note that (i) of Definition 1.3 is equivalent to the following property: there exists a function $\bar{f} : \mathbb{R}^n \to \mathbb{R}$ which is C^k around x and which agrees with f on $\mathcal{M}$ near x.

In particular, a partly smooth function is smooth when restricted to the manifold $\mathcal{M}$. This property will be useful to define Newton's method on $\mathcal{M}$ (see Algorithm 1.8). Other properties of this definition are crucial to explain algorithms related to the $\mathcal{U}$-Lagrangian, especially for the key Theorem 2.12.

Assumptions. The following assumptions are made throughout the paper.

Assumption 1.4 (Convexity). *We consider a function f that is convex and real-valued over all of $\mathbb{R}^n$.*

This assumption is made to accommodate the $\mathcal{U}$-Newton method which does not apply to non-convex functions. Note that Clarke regularity (point (ii) of Definition 1.3) is implied by convexity.

Assumption 1.5 (Partial smoothness). *There exists a C^∞-manifold $\mathcal{M}$ containing a minimizer of f, such that f is C^∞-partly smooth relative to $\mathcal{M}$ at all $x \in \mathcal{M}$.*

The underlying smoothness has been chosen to be C^∞ because it is a convenient assumption for manipulating geometric objects. Most of the results can be obtained with f only C^2-partly smooth on a C^2-sub-manifold $\mathcal{M}$. In the rest of the paper, the term "smooth" for functions or manifolds stands for "C^∞".

Since the $\mathcal{U}$-Lagrangian is a principal object of study in this paper, we limit the development to the case of convex f. Nevertheless, much of the second-order theory may be applied to nonconvex functions with a smooth substructure. The notion of partial smoothness applies to the nonconvex case, as does the primal-dual gradient (pdg) structure of Mifflin and Sagastizábal [16, 18, 19]. Indeed, the relationships among the $\mathcal{U}$-Lagrangian, $\mathcal{VU}$-decomposition, partial smoothness and pdg structure are numerous, but they are largely beyond the scope of this paper. We will only hint at a few connections; see [8, 12, 16, 18, 19] for thorough treatments.

Examples. We give below two examples of our situation. The simple first example is used as an illustration in the beginning of Section 3.

Example 1.6 (Basic example). Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x) := \max\{x_1,\ \|x-(1,0)\|^2-1\}.$$

Its minimum value is 0 and is achieved at (0, 0) belonging to

$$\mathcal{M} = \{x \mid x_1 = \|x-(1,0)\|^2-1\}$$

which is a smooth sub-manifold of $\mathbb{R}^2$. Let us prove that f is partly smooth around (0, 0) relative to $\mathcal{M}$. An equation of $\mathcal{M}$ is ${x_1}^2 - 3x_1 + {x_2}^2 = 0$. The restriction of f to $\mathcal{M}$ is simply $(x_1, x_2) \in \mathcal{M} \mapsto x_1$, which is smooth. Observe that $\partial f(x)$ is the segment joining (1, 0) and $(2x_1-2, 2x_2)$, so the set-valued map ∂f restricted to $\mathcal{M}$ is continuous around (0, 0), and

$$\mathrm{T}_{\mathcal{M}}(0,0) = 0 \oplus \mathbb{R}, \qquad \mathrm{N}_{\mathcal{M}}(0,0) = \mathbb{R} \oplus 0.$$

Since $f(t,0) = \max\{t, t^2-2t\}$ is not differentiable at 0, we can conclude that f is partly smooth relative to $\mathcal{M}$ around (0, 0).

Example 1.7 (Maximum eigenvalue function). Let $\mathcal{S}_m$ be the Euclidean space of symmetric m by m matrices. We denote by $\lambda_1(X) \geq \cdots \geq \lambda_m(X)$ the eigenvalues of $X \in \mathcal{S}_m$.

The set $\mathcal{M}_r = \{A \in \mathcal{S}_m,\ \lambda_1(A) = \cdots = \lambda_r(A) > \lambda_{r+1}(A)\}$ of symmetric matrices whose maximum eigenvalue has a given multiplicity r is a smooth sub-manifold of $\mathcal{S}_m$ (see [21] for example). Furthermore, the maximum eigenvalue function λ_1 is partly smooth at $X \in \mathcal{M}_r$ relative to $\mathcal{M}_r$ (see [12]).

Considering a smooth function $F : \mathbb{R}^n \to \mathcal{S}_m$, the chain rule of [12, Th. 4.2] then implies that $\lambda_1 \circ F$ is partly smooth at $z \in F^{-1}(\mathcal{M}_r)$ relative to $F^{-1}(\mathcal{M}_r)$ if

$$\ker \mathrm{D}F(z)^* \cap \mathrm{N}_{\mathcal{M}_r}(F(z)) = \{0\}.$$

In the particular case where F is affine, we recover the transversal assumption of [21, Def. 5.4].

1.3. Sequential Newton method

In this paper, we consider an iteration that is not exactly a Newton method, but rather a "sequential Newton method" as introduced in [25]. The function to which the Newton idea is applied changes at every iteration.

Algorithm 1.8 (Sequential Newton). *Given $x \in \mathcal{M}$ and $\{\varphi_x\}_x$ a parameterization family of $\mathcal{M}$, repeat the update $x \leftarrow N(x)$ where*

$$h(x) = -\left[\nabla^2(\bar{f} \circ \varphi_x)(0)\right]^{-1} \nabla(\bar{f} \circ \varphi_x)(0) \tag{1.3}$$

$$N(x) = \varphi_x(h(x)). \tag{1.4}$$

We call $N(x)$ the Newton update*, $N(x) - x$ the* Newton step*, and $h(x)$ the* Newton direction.

We will consider applying the Newton update $N(x)$ at an arbitrary point $x \in \mathcal{M}$. Issues of global convergence—and even the problem of existence of the full Newton step—are ignored in this paper. It is not our intent to describe a practical algorithm, but instead to explore the connections between various formulations of Newton methods when they are well-defined.

The $\mathcal{U}$-Newton and Riemannian Newton methods developed in subsequent sections take the form of Algorithm 1.8 to which the standard proof of local convergence of Newton methods does not apply. We will prove quadratic convergence for each of the sequential Newton methods we consider by comparing with a prototype Newton iteration that is intrinsically defined on the manifold (Lemma 4.3).

2. $\mathcal{VU}$-Theory for Partly Smooth Convex Functions

The general $\mathcal{VU}$-theory for the study of the non-smooth behavior of convex functions was introduced in [11]. We study here what is brought by our particular context: how do the notions of $\mathcal{VU}$-decomposition, $\mathcal{U}$-Lagrangian and fast tracks behave under Assumption 1.5? We also develop some continuity properties of the gradient of the $\mathcal{U}$-Lagrangian that will be useful later.

2.1. $\mathcal{VU}$-decomposition

The idea is to decompose $\mathbb{R}^n$ into two orthogonal subspaces $\mathcal{U}(x)$ and $\mathcal{V}(x)$, such that $\mathcal{V}(x)$ contains the non-smooth behavior of f at x. For $x \in \mathbb{R}^n$ and an arbitrary $g \in \partial f(x)$, we define the following subspaces of $\mathbb{R}^n$:

$$\mathcal{V}(x) := \operatorname{lin}\left(\partial f(x) - g\right), \qquad \mathcal{U}(x) := \mathcal{V}(x)^{\perp}.$$

Notice that $\mathcal{U}(x)$ and $\mathcal{V}(x)$ are actually independent of the choice of g.

Near x, the function f appears smooth in directions from the subspace $\mathcal{U}$, while $\mathcal{V}$ determines the directions of nonsmoothness (see [11, Def. 2.1 and Prop. 2.2]). Actually, when $x \in \mathcal{M}$, the subspaces $\mathcal{U}$ and $\mathcal{V}$ are respectively the tangent and normal subspaces of the manifold $\mathcal{M}$ at x.

Lemma 2.1 (Interpretation of $\mathcal{U}$ and $\mathcal{V}$). *Let Φ define a local equation of $\mathcal{M}$ around $x \in \mathcal{M}$. The following relations hold:*

$$\begin{aligned} \mathcal{U}(x) &= \mathrm{T}_{\mathcal{M}}(x) = \ker(\mathrm{D}\Phi(x)) \\ \mathcal{V}(x) &= \mathrm{N}_{\mathcal{M}}(x) = \mathrm{range}(\mathrm{D}\Phi(x)^*). \end{aligned}$$

Proof. The subspace $\mathcal{U}(x)$ is exactly the space of directions for which the directional derivative of f is linear [11, Prop. 2.2(ii)], and so the normal sharpness of f relative to $\mathcal{M}$ (Assumption 1.5(iii)) is equivalent to $\mathrm{T}_{\mathcal{M}}(x) = \mathcal{U}(x)$ [12, Note 2.9(a)]. The remaining equalities follow directly from the local equations and the definition of $\mathcal{V}(x)$. □

Theorem 2.2. *Let $\bar{x} \in \mathcal{M}$. Then there is a neighborhood Ω of $(\bar{x}, 0)$ in $\mathrm{T}\mathcal{M}$ and a unique smooth function $v : \Omega \to \mathbb{R}^n$ such that for $x \in \mathcal{M}$ close to $\bar{x}$ and $d \in \mathbb{R}^n$ small enough satisfying $x + d \in \mathcal{M}$, we have*

$$d = u + v(x, u)$$

with $u = \mathrm{P}_{\mathcal{U}(x)}(d)$ and $v(x, u) \in \mathcal{V}(x)$. In addition, for $u \in \mathcal{U}(x)$

$$\|v(x, u)\| = O(\|u\|^2).$$

Proof. Let Φ be a local equation of $\mathcal{M}$ around $\bar{x}$. Consider Ψ defined by

$$\Psi(x, u, v) = \Phi(x + u + v)$$

for $x \in \mathcal{M}$ close to $\bar{x}$, $u \in \mathcal{U}(x) = \mathrm{T}_{\mathcal{M}}(x)$ and $v \in \mathcal{V}(x) = \mathrm{N}_{\mathcal{M}}(x)$. The partial differential $\mathrm{D}_v\Psi(\bar{x}, 0, 0)$ is, for $v \in \mathrm{N}_{\mathcal{M}}(\bar{x})$,

$$\mathrm{D}_v\Psi(\bar{x}, 0, 0)v = \mathrm{D}\Phi(\bar{x})v.$$

Thus $\mathrm{D}_v\Psi(\bar{x}, 0, 0)$ is surjective from $\mathcal{V}(\bar{x})$ to $\mathbb{R}^{n-p}$, so it is a bijection. The implicit function theorem yields that there exists a unique smooth function $v(x, u)$ such that for all $x \in \mathcal{M}$ close to $\bar{x}$, all u close to 0 and all v close to 0,

$$\Psi(x, u, v) = 0 \iff v = v(x, u),$$

which means

$$x + d \in \mathcal{M} \iff d = u + v(x, u)$$

with $u = \mathrm{P}_{\mathcal{U}(x)}(d)$. At $d = 0$, there holds $v(x, 0) = 0$. Compute the partial derivative of v at $(x, 0)$,

$$\mathrm{D}_u v(x, 0) = -[\mathrm{D}_v\Psi(x, 0, 0)]^{-1}[\mathrm{D}_u\Psi(x, 0, 0)].$$

The inverse of $\mathrm{D}_v\Psi(x, 0, 0)$ exists by continuity since $\mathrm{D}_v\Psi(\bar{x}, 0, 0)$ is a bijection, and $\mathrm{D}_u\Psi(x, 0, 0) = 0$ since $\mathcal{U}(x) = \ker(\mathrm{D}\Phi(x)) = \ker(\mathrm{D}\Psi(x, 0, 0))$. Therefore $\mathrm{D}_u v(x, 0) = 0$, which implies that $\|v(x, u)\| = O(\|u\|^2)$. □

The $\mathcal{VU}$-decomposition of the space will induce via v the tangential parameterization of $\mathcal{M}$ (see Lemma 3.3). Particularizing Theorem 2.2 for $x = \bar{x}$, we obtain the following statement, which is part of [12, Th. 6.1].

Corollary 2.3 (Manifold as a graph). *Let $x \in \mathcal{M}$. Then there is a neighborhood Θ of 0 in $\mathcal{U}(x)$ and a unique smooth function $v : \Theta \to \mathcal{V}(x)$ such that for $d \in \mathbb{R}^n$ small enough satisfying $x + d \in \mathcal{M}$, we have*

$$d = u + v(u)$$

with $u = \mathrm{P}_{\mathcal{U}(x)}(d)$. Furthermore,

$$v(0) = 0 \quad \text{and} \quad \mathrm{D}v(0) = 0,$$

and then for $u \in \mathcal{U}(x)$

$$\|v(u)\| = O(\|u\|^2).$$

At this point we remind the reader that x and $\bar{x}$ are not meant to be fixed at any value; they are free variables, ranging over $\mathcal{M}$ unless otherwise specified. We use $\bar{x}$ to denote an arbitrary point with some property, and then x is an arbitrary nearby point. Nevertheless, for brevity we often drop the explicit dependence of $\mathcal{U}$, $\mathcal{V}$ and v on x when it is clear which x is intended.

2.2. $\mathcal{U}$-Lagrangian

Given $g \in \partial f(x)$, the *$\mathcal{U}$-Lagrangian* of f at x [11] is the function $L^g_{\mathcal{U}} : \mathcal{U}(x) \to \mathbb{R}$ defined by

$$L^g_{\mathcal{U}}(u) := \inf_{v \in \mathcal{V}(x)} \left\{ f(x + u + v) - g^\top v \right\}. \tag{2.1}$$

The $\mathcal{U}$-Lagrangian an effective way to extract the smooth behavior of f along $\mathcal{U}(x)$. Theorem 3.3 in [11] says that $L^g_{\mathcal{U}}$ is differentiable at $u = 0$, and that its derivative at 0 is given by

$$\nabla L^g_{\mathcal{U}}(0) = \mathrm{P}_{\mathcal{U}}(g).$$

For our purposes, it is important to emphasize that the derivative at 0 is actually independent of $g \in \partial f(x)$. We call it the *$\mathcal{U}$-gradient* of f at x.

Lemma 2.4 ($\mathcal{U}$-gradient). *All $g \in \partial f(x)$ have the same projection on $\mathcal{U}(x)$, which we denote by $g_{\mathcal{U}}(x)$. Moreover,*

$$g_{\mathcal{U}}(x) := \mathrm{P}_{\mathcal{U}}(g) = \mathrm{P}_{\mathrm{aff}\, \partial f(x)}(0).$$

Proof. Set $h := \mathrm{P}_{\mathrm{aff}\, \partial f(x)}(0)$, and notice that $\mathcal{V} = \mathrm{lin}(\partial f(x) - h)$. Since $\mathcal{U}$ and aff $\partial f(x)$ are perpendicular, h lies in $\mathcal{U}$. Given $g \in \partial f(x)$,

$$\mathrm{P}_{\mathcal{U}}(g) = \mathrm{P}_{\mathcal{U}}(g - h) + \mathrm{P}_{\mathcal{U}}(h) = 0 + \mathrm{P}_{\mathcal{U}}(h) = h,$$

which ends the proof. □

Note that the $\mathcal{U}$-gradient $g_{\mathcal{U}}(x)$ may not be in $\partial f(x)$ in general. However, it is the case if x is close to a point $\bar{x}$ where $g_{\mathcal{U}}(\bar{x}) \in \mathrm{ri}\, \partial f(\bar{x})$ holds (see Theorem 2.12 below).

2.3. $\mathcal{U}$-Hessian

The set of minimizers in (2.1) is denoted by $W^g(u)$:

$$W^g(u) := \operatorname*{argmin}_{w \in \mathcal{V}(x)} \left\{ f(x+u+w) - g^\top w \right\}. \tag{2.2}$$

If g actually belongs to $\operatorname{ri} \partial f(x)$, then $W^g(u)$ is non-empty for all $u \in \mathcal{U}(x)$ [11, Th. 3.2]. Furthermore the partial smoothness assumption implies that a minimizing w is unique for each u small enough, and it coincides with $v(u)$.

Theorem 2.5 (Unique winner). *Let $g \in \operatorname{ri} \partial f(x)$. For all $u \in \mathcal{U}(x)$ small enough, $W^g(u)$ reduces to $\{v(u)\}$ (given by Corollary 2.3). Moreover $L^g_{\mathcal{U}}$ is smooth in a neighborhood of* 0 *in* $\mathcal{U}(x)$.

Proof. $W^g(u)$ reduces to $v(u)$ thanks to partial smoothness (Theorem 6.1 in [12]). It follows that for all $u \in \mathcal{U}$ small enough,

$$L^g_{\mathcal{U}}(u) = f(x+u+v(u)) - g^\top v(u). \tag{2.3}$$

Corollary 2.3 and Assumption 1.5 yield that $L^g_{\mathcal{U}}$ is smooth around 0. □

Theorem 2.5 shows that the $\mathcal{U}$-Lagrangian is useful for the analytic construction of the implicitly-defined v. Moreover, it ensures the existence of $\nabla^2 L^g_{\mathcal{U}}(0)$, the so-called *$\mathcal{U}$-Hessian* of f at x [11]. A second-order-like expansion of f on $\mathcal{M}$ is obtained from the $\mathcal{U}$-Hessian.

Theorem 2.6 ("Second-order" expansion). *Let $g \in \operatorname{ri} \partial f(x)$. For $d \in \mathbb{R}^n$ such that $x+d \in \mathcal{M}$ we have*

$$f(x+d) = f(x) + g^\top d + \tfrac{1}{2}(\mathrm{P}_{\mathcal{U}} d)^\top \left[\nabla^2 L^g_{\mathcal{U}}(0)\right] \mathrm{P}_{\mathcal{U}} d + O(\|d\|^3). \tag{2.4}$$

Proof. This is Theorem 3.9 in [11]. There the remainder term is $o(\|d\|^2)$, but the extra smoothness of $L^g_{\mathcal{U}}$ actually makes it $O(\|d\|^3)$. □

Example 2.7 (Maximum eigenvalue function). Since the transversality assumption yields that $\lambda_1 \circ A$ is partly smooth relative to $A^{-1}(\mathcal{M}_r)$, Theorem 5.9 of [21] is a particular case of Theorem 2.5. Moreover, Corollary 4.13 in [21] corresponds to Theorem 2.6 written for $f = \lambda_1$.

2.4. Fast tracks

In this section only, we drop the partial smoothness assumption on f for the purpose of establishing connections between fast tracks and partial smoothness. The definition below is essentially extracted from [17], where the concept of fast track was first introduced in order to define Newton-like methods for a class of nonsmooth convex functions.

Definition 2.8 (Fast track). *Let $x \in \mathbb{R}^n$. We say that $x+u+w(u)$ is a* fast track *leading to x if for all u small enough*

(i) $w : \mathcal{U} \to \mathcal{V}$ *is a smooth function such that* $w(u) \in W^g(u)$ *for all* $g \in \operatorname{ri} \partial f(x)$*;*
(ii) $u \mapsto f(x + u + w(u))$ *is a smooth function.*

For consistency of the paper, this definition assumes that w is C^∞ (smooth) instead of only C^2 as in [17]. Nevertheless, the proof of the following theorem can be easily adapted to the C^2 case. Note however that we slightly extend the definition: x is not necessarily a minimizer of f (as in [17, Def. 2.1]).

Theorem 2.9 (Fast track and partial smoothness). *Let* $x \in \mathbb{R}^n$*. Suppose that* $x + u + w(u)$ *is a fast track leading to* x*, and define* $\psi : \mathcal{U}(x) \to \mathbb{R}^n$ *by* $\psi(u) = x + u + w(u)$*. Then* f *is partly smooth at* x *relative to* $\mathcal{M} = \{\psi(u) \mid u \in \mathcal{U}(x)\}$ *which is a manifold. Furthermore* $w(u)$ *is the function* $v(u)$ *of Corollary 2.3.*

Proof. Since w is smooth, so is ψ. We compute $\mathrm{D}\psi(0) = I + \mathrm{D}w(0)$. Since $w(u) \in W^g(u)$ for any $g \in \operatorname{ri} \partial f(x)$, Corollary 3.5 of [11] yields that

$$\|w(u)\| = o(\|u\|), \tag{2.5}$$

and then $\mathrm{D}w(0) = 0$. Thus $\mathrm{D}\psi(0) = I$ is a bijection, which implies by the local inverse theorem that ψ is a smooth diffeomorphism. We can conclude that $\mathcal{M}$ is a smooth submanifold, and $\operatorname{range}(\mathrm{D}\psi(0)) = \mathcal{U}$ is the tangent space at x. Finally, the uniqueness in Corollary 2.3 yields that w is the function v of $\mathcal{M}$.

Let us check the four points of Definition 1.5. First, f is smooth on $\mathcal{M}$ by Definition 2.8(ii). Second, f is convex hence Clarke regular [24]. Third, since $\mathcal{U} = \mathrm{T}_{\mathcal{M}}(x)$, we have normal sharpness by Definition 2.1 of [11]. The only point to prove is the inner semi-continuity of the restriction of ∂f to $\mathcal{M}$, since ∂f is already outer semi-continuous [9, VI.6.2.4].

Definition 2.8 implies: for all $g \in \operatorname{ri} \partial f(x)$, the function $u \mapsto L^g_{\mathcal{U}}(u)$ is also smooth at all u in a neighborhood Ω of 0. Theorem 3.3 of [11] says that

$$\partial L^g_{\mathcal{U}}(u) = \big\{h \;\big|\; h + \mathrm{P}_{\mathcal{V}}(g) \in \partial f(\psi(u))\big\}.$$

In our case, this means that for all $g \in \operatorname{ri} \partial f(x)$ and all $u \in \Omega$,

$$\nabla L^g_{\mathcal{U}}(u) + \mathrm{P}_{\mathcal{V}}(g) \in \partial f(\psi(u)),$$

which can be rewritten with Lemma 2.4 as

$$\nabla L^g_{\mathcal{U}}(u) - g_{\mathcal{U}} + g \in \partial f(\psi(u)). \tag{2.6}$$

Furthermore, note that for all g and g' in $\operatorname{ri} \partial f(x)$,

$$L^g_{\mathcal{U}}(u) - L^{g'}_{\mathcal{U}}(u) = (g' - g)^\top w(u)$$

for u small enough, by Definition 2.8(i).

Let $\varepsilon > 0$ be arbitrary. The boundedness of $\partial f(x)$ and (2.5) imply that there exists a neighborhood N of 0 in $\mathcal{U}$ such that

$$\forall u \in N, \quad \forall g, g' \in \operatorname{ri} \partial f(x), \qquad \|\nabla L^g_{\mathcal{U}}(u) - \nabla L^{g'}_{\mathcal{U}}(u)\| \le \varepsilon. \tag{2.7}$$

Fix $g \in \operatorname{ri} \partial f(x)$, and use (2.6) and (2.7) to get

$$\nabla L^g_{\mathcal{U}}(u) - g_{\mathcal{U}} + \operatorname{ri} \partial f(x) \subset \partial f(\psi(u)) + B(0, \varepsilon).$$

Since $\partial f(x)$ is closed for all x, we obtain

$$\nabla L^g_{\mathcal{U}}(u) - g_{\mathcal{U}} + \partial f(x) \subset \partial f(\psi(u)) + B(0, \varepsilon).$$

Observe that $\nabla L^g_{\mathcal{U}}(u)$ tends to $\nabla L^g_{\mathcal{U}}(0) = g_{\mathcal{U}}$ when u tends to 0. Then restricting N if necessary, we have $\|\nabla L^g_{\mathcal{U}}(u) - g_{\mathcal{U}}\| \leq \varepsilon$ and then

$$\partial f(\psi(0)) = \partial f(x) \subset \partial f(\psi(u)) + B(0, 2\varepsilon).$$

This expresses the inner semi-continuity of ∂f on $\mathcal{M}$. We can conclude that f is partly smooth on $\mathcal{M}$, which ends the proof. □

Corollary 2.10. *There is a fast track leading to x if and only if f is partly smooth at x.*

Proof. If f is partly smooth at x relative to $\mathcal{M}$ then v provided by Corollary 2.3 determines a fast track. The other direction is simply Theorem 2.9. □

The same result appears in [8, Th. 3.1], which the authors discovered after this paper was written.

2.5. Continuity properties of $\mathcal{U}$-gradient

In this subsection, we prove two properties of $g_{\mathcal{U}}(\cdot)$, namely its continuity and the persistence of the property $g_{\mathcal{U}}(x) \in \operatorname{ri} \partial f(x)$ for small perturbations of x.

Lemma 2.11 (Continuity of $g_{\mathcal{U}}$). *The function $g_{\mathcal{U}} : \mathcal{M} \to \mathbb{R}^n$ is continuous on $\mathcal{M}$.*

Proof. Recall from Lemma 2.4 that

$$g_{\mathcal{U}}(x) = \mathrm{P}_{\mathcal{U}(x)}(\partial f(x)).$$

Let $\bar{x} \in \mathcal{M}$ and Φ be a local equation of $\mathcal{M}$ around $\bar{x}$. Subdifferential continuity (Definition 1.3(iv)) means that $x \mapsto \partial f(x)$ is continuous on $\mathcal{M}$. It is easy to see that $x \mapsto \mathrm{P}_{\mathcal{U}(x)}$ is also continuous for x near $\bar{x}$: since $\mathcal{U}(x) = \ker \mathrm{D}\Phi(x)$ (Lemma 2.1), an expression for $\mathrm{P}_{\mathcal{U}(x)}$ is

$$\mathrm{P}_{\mathcal{U}(x)} = I - \mathrm{D}\Phi(x)^*[\mathrm{D}\Phi(x)\mathrm{D}\Phi(x)^*]^{-1}\mathrm{D}\Phi(x).$$

We can then conclude that $g_{\mathcal{U}}$ is continuous around $\bar{x}$. □

Theorem 2.12 (Persistence of $g_{\mathcal{U}}$ as an interior subgradient). *Let $\bar{x} \in \mathcal{M}$. If $g_{\mathcal{U}}(\bar{x}) \in \operatorname{ri} \partial f(\bar{x})$, then $g_{\mathcal{U}}(x) \in \operatorname{ri} \partial f(x)$ for any $x \in \mathcal{M}$ close enough to $\bar{x}$.*

Proof. Observe first that

$$\mathcal{V}(x) = \operatorname{lin}(\partial f(x) - g_{\mathcal{U}}(x)).$$

Notice from Lemma 2.4 that the dimension of $\mathcal{V}(x) = \mathrm{N}_{\mathcal{M}}(x)$ is constant (equal to $n - p$). We also deduce that there is a basis of $\mathcal{V}(x)$ depending continuously on $x \in \mathcal{M}$ (the columns of a matrix representing $\mathrm{D}\Phi(x)^*$). With this basis, it is easy to construct a continuous function $x \mapsto \psi_x$ such that

$$\psi_x : \mathcal{V}(x) \longrightarrow \mathbb{R}^{n-p}$$

is a linear bijection between $\mathcal{V}(x)$ and $\mathbb{R}^{n-p}$. Consider then $C : \mathcal{M} \rightrightarrows \mathbb{R}^{n-p}$ defined by

$$C(x) = \psi_x(\partial f(x) - g_{\mathcal{U}}(x)).$$

Continuity of ∂f (by partial smoothness assumption (iv)), of $g_{\mathcal{U}}$ (by Lemma 2.11) and of ψ_x (by construction) yield the continuity of C on $\mathcal{M}$ around $\bar{x}$. Furthermore, observe that

$$g_{\mathcal{U}}(x) \in \operatorname{ri} \partial f(x) \quad \Leftrightarrow \quad 0 \in \operatorname{int} C(x).$$

To prove this, consider $r > 0$ such that

$$0 \in B(0, r) \cap \mathcal{V}(x) \subset \partial f(x) - g_{\mathcal{U}}(x),$$

and observe that

$$0 = \psi_x(0) \in B(0, r/\|\psi_x^{-1}\|) \subset C(x)$$

since ψ_x is a linear bijection.

Now, suppose for contradiction that there exists a sequence $\{x_k\}$ of points in $\mathcal{M}$ such that x_k tends to x and $g_{\mathcal{U}}(x_k) \notin \operatorname{ri} \partial f(x_k)$. Set $C_k = C(x_k)$ so that $0 \notin \operatorname{int} C_k$. For all k, C_k is convex since ψ_{x_k} is linear. Now separate 0 from $\operatorname{int} C_k$: there exist $s_k \in \mathbb{R}^{n-p}$ with $\|s_k\| = 1$ such that

$$\forall k \in \mathbb{N}, \quad \forall y \in C_k, \qquad s_k^\top y \leq 0. \tag{2.8}$$

Extracting a subsequence if necessary, we can suppose that $s_k \to s$ with $\|s\| = 1$. Since $0 \in \operatorname{int} C(\bar{x})$, let $r > 0$ be such that $B(0, r) \subset C(\bar{x})$. Let $v \in B(0, r)$; the continuity of C implies that there are $v_k \in C_k$ such that $v_k \to v$. With (2.8), we can write

$$\forall k \in \mathbb{N}, \qquad s_k^\top v_k \leq 0.$$

Passing to the limit, this gives $s^\top v \leq 0$. This can be done for any $v \in B(0, r)$, so we have $s^\top v = 0$ for all $v \in B(0, r)$. We conclude that $s = 0$, which contradicts $\|s\| = 1$. □

Remark 2.13. In the particular case $f = \lambda_1$, the last theorem corresponds to [21, Prop. 6.9(iii)]. A similar result for pdg-structured functions can be recovered from [19, Th. 4.2] , where the subgradient is shown to be C^1.

Theorem 2.12 leads to another interpretation of $g_{\mathcal{U}}$ in a neighborhood of a "sharp" minimizer of f.

Corollary 2.14. *Let $\bar{x}$ be a minimizer of f such that* $0 \in \operatorname{ri} \partial f(\bar{x})$. *Then for all x in a neighborhood of $\bar{x}$ in $\mathcal{M}$,*

$$\mathrm{P}_{\partial f(x)}(0) = g_{\mathcal{U}}(x).$$

Proof. Since $0 = g_{\mathcal{U}}(\bar{x}) \in \operatorname{ri} \partial f(\bar{x})$, Theorem 2.12 yields that $g_{\mathcal{U}}(x) \in \operatorname{ri} \partial f(x)$ for all x in a neighborhood of $\bar{x}$ in $\mathcal{M}$. Using Lemma 2.4 we have $\mathrm{P}_{\operatorname{aff} \partial f(x)}(0) = g_{\mathcal{U}}(x) \in \operatorname{ri} \partial f(x)$, and thus $g_{\mathcal{U}} = \mathrm{P}_{\partial f(x)}(0)$. □

3. The $\mathcal{U}$-Newton Method

3.1. Curvature in the tangential parameterization

A difficulty of interpretation arises in Theorem 2.6. The $\mathcal{VU}$-decomposition, the manifold $\mathcal{M}$ and the tangential parameterization $u \mapsto x + u + v(u)$ are all geometric properties of f at x. The second-order behavior of f along a tangentially parameterized curve should not depend on the choice of g used to define the $\mathcal{U}$-Lagrangian. Indeed, Lemma 2.4 shows that the $\mathcal{U}$-gradient $\nabla L^g_{\mathcal{U}}(0)$ is independent of g. But the following example demonstrates that the $\mathcal{U}$-Hessian $\nabla^2 L^g_{\mathcal{U}}(0)$ depends on g. Other similar examples can be found in [15].

Example 3.1 (Dependence on the subgradient). Consider the function f of Example 1.6. Its subdifferential at $(0, 0)$ is $\partial f(0, 0) = [-2, 1] \times \{0\}$, so we obtain $\mathcal{U} = 0 \oplus \mathbb{R}$ and $\mathcal{V} = \mathbb{R} \oplus 0$. Since $x_1 = \frac{3}{2} - \sqrt{\frac{9}{4} - x_2{}^2}$ is a local equation around $(0, 0)$ of $\mathcal{M}$, there is, for $u = (0, u_2) \in \mathcal{U}$,

$$v(u) = \begin{bmatrix} \frac{3}{2} - \sqrt{\frac{9}{4} - u_2{}^2} \\ 0 \end{bmatrix},$$

whose derivative is

$$\mathrm{D}v(u) = \begin{bmatrix} 0 & \dfrac{2u_2}{\sqrt{9 - 4u_2{}^2}} \\ 0 & 0 \end{bmatrix}.$$

Choose an arbitrary $g \in \partial f(x)$, so $g = (\gamma, 0)$ for some $\gamma \in [-2, 1]$. For any $u = (0, u_2) \in \mathcal{U}$, the $\mathcal{U}$-Lagrangian is

$$L^g_{\mathcal{U}}(u) = (1 - \gamma)\left(\frac{3}{2} - \sqrt{\frac{9}{4} - u_2{}^2}\right),$$

and its derivatives are

$$\nabla L^g_{\mathcal{U}}(u) = \begin{bmatrix} 0 \\ (1-\gamma)\dfrac{2u_2}{\sqrt{9 - 4u_2{}^2}} \end{bmatrix} \quad \text{and} \quad \nabla^2 L^g_{\mathcal{U}}(u) = \begin{bmatrix} 0 & 0 \\ 0 & (1-\gamma)\dfrac{18}{(9 - 4u_2{}^2)^{3/2}} \end{bmatrix}.$$

We see that $\nabla L^g_{\mathcal{U}}(0, 0) = 0$, but $\nabla^2 L^g_{\mathcal{U}}(0, 0)$ depends on g. □

Why, then, does the $\mathcal{U}$-Hessian—which seems to determine the curvature of f in (2.4)—depend on g? The answer is that the $\mathcal{U}$-Hessian does not entirely determine the curvature of f because the "linear" term $g^\top d$ may have curvature! Since the trajectory of d in Theorem 2.6 is constrained by $x + d \in \mathcal{M}$, the term $g^\top v(u)$ contributes to the curvature of f.

To see this, take a trajectory on $\mathcal{M}$: fix $u \in \mathcal{U}$ and substitute $d(t) = tu + v(tu)$ into the expansion (2.4) to get

$$f(x + d(t)) = f(x) + g^\top(tu + v(tu)) + \tfrac{1}{2}t^2u^\top[\nabla^2 L^g_{\mathcal{U}}(0)]u + O(t^3). \tag{3.1}$$

By Corollary 2.3, we can thus write

$$v(tu) = \tfrac{1}{2}t^2[\mathrm{H}v(0)](u, u) + O(t^3),$$

where $[\mathrm{H}v(0)]$ denotes the Hessian of v at the point 0. More generally, throughout the paper, if Ψ is a twice differentiable function between two vector spaces X and Y, then $\mathrm{H}\Psi(x)$ is its Hessian at $x \in X$ which is a bilinear mapping from $X \times X$ to Y, and $[\mathrm{H}\Psi(x)](\delta_1, \delta_2)$ denotes its value at $(\delta_1, \delta_2) \in X \times X$. Now (3.1) becomes the second-order Taylor expansion

$$f(x + d(t)) = f(x) + tg^\top u + \tfrac{1}{2}t^2\big(g^\top[\mathrm{H}v(0)](u, u) + u^\top[\nabla^2 L^g_{\mathcal{U}}(0)]u\big) + O(t^3), \tag{3.2}$$

and the second derivative includes the extra term $g^\top[\mathrm{H}v(0)](u, u)$. The next lemma shows that $\mathrm{H}v(0)$ cannot in general be ignored.

Lemma 3.2. *Let $x \in \mathcal{M}$ and let Φ define a local equation of $\mathcal{M}$. If the restriction of* $\mathrm{H}\Phi(x)$ *to $\mathcal{U}(x) \times \mathcal{U}(x)$ is not identically null, then* $\mathrm{H}v(0)$ *is not identically null either.*

Proof. We have $\Phi(x + u + v(u)) = 0$ for u small enough. From Corollary 2.3, $u \mapsto \Phi(x + u + v(u))$ is smooth around $u = 0$. Differentiating the equation once around $u = 0$ we get, for all $\delta_1 \in \mathcal{U}$,

$$\mathrm{D}\Phi(x + u + v(u))(I + \mathrm{D}v(u))\delta_1 = 0,$$

and differentiating again we get, for all $\delta_1, \delta_2 \in \mathcal{U}$,

$$[\mathrm{H}\Phi(x + u + v(u))]\Big((I + \mathrm{D}v(u))\delta_1,\ (I + \mathrm{D}v(u))\delta_2\Big) + \mathrm{D}\Phi(x + u + v(u))\Big([\mathrm{H}v(u)](\delta_1, \delta_2)\Big) = 0.$$

At $u = 0$ we have $\mathrm{D}v(0) = 0$ so for all $\delta_1, \delta_2 \in \mathcal{U}$,

$$[\mathrm{H}\Phi(x)](\delta_1, \delta_2) + \mathrm{D}\Phi(x)\Big([\mathrm{H}v(0)](\delta_1, \delta_2)\Big) = 0. \tag{3.3}$$

By hypothesis, there exist $\bar{\delta}_1, \bar{\delta}_2 \in \mathcal{U}$ such that $[\mathrm{H}\Phi(x)](\bar{\delta}_1, \bar{\delta}_2) \neq 0$, hence $[\mathrm{H}v(0)](\bar{\delta}_1, \bar{\delta}_2) \neq 0$. □

Therefore, use of the $\mathcal{U}$-Hessian $\nabla^2 L^g_{\mathcal{U}}(0)$ alone to model the curvature of f on $\mathcal{M}$ may not lead to a true Newton method. However, if g can be chosen in $\mathcal{U}$ (so that $g^\top v(u) = 0$ for all $u \in \mathcal{U}$), then the $\mathcal{U}$-Hessian does accurately model the curvature of f. Observe from Lemma 2.4 that if $\partial f(x) \cap \mathcal{U}$ is non-empty, then it is equal to $\{g_{\mathcal{U}}\}$. The $\mathcal{U}$-gradient $g_{\mathcal{U}}$ is thus the only possible choice for such a g. If $g_{\mathcal{U}} \notin \partial f(x)$, no $\mathcal{U}$-Hessian can give the correct curvature of f. Notice that, if x is a sharp minimizer of f, then $g_{\mathcal{U}} = 0 \in \operatorname{ri} \partial f(x)$ and the $\mathcal{U}$-Hessian with $g = 0$ precisely models the curvature of f around x. But we show in the next section that a proper $\mathcal{U}$-Newton method is defined using $g_{\mathcal{U}}$, even if $g_{\mathcal{U}} \notin \partial f(x)$.

3.2. The $\mathcal{U}$-Newton method

We now describe a sequential Newton method based on the $\mathcal{U}$-Lagrangian theory. Following Algorithm 1.8, we need a parameterization family of $\mathcal{M}$. The function $v(x, u)$ of Theorem 2.2 provides the parameterization

$$\varphi_x^{\tan}(u) := x + u + v(x, u), \tag{3.4}$$

for $u \in \mathcal{U}(x)$ small enough. The superscript "tan" stands for "tangential".

Lemma 3.3 (Tangential parameterization). *The function $\varphi_x^{\tan}$ is a local parameterization of $\mathcal{M}$ around x, and the family $\{\varphi_x^{\tan}\}_x$ is a smooth parameterization family.*

Proof. Straightforward from Theorem 2.2 (note that $[\varphi_x^{\tan}]^{-1} = \mathrm{P}_{\mathcal{U}(x)}(\cdot - x)$). □

By *$\mathcal{U}$-Newton method* we mean the sequential Newton method of Algorithm 1.8 using the parameterization family $\{\varphi_x^{\tan}\}_x$. To define the algorithm, we need the gradient and Hessian of $f \circ \varphi_x^{\tan}$, or equivalently of $\bar f \circ \varphi_x^{\tan}$ with $\bar f$ given by Assumption 1.5, at $u = 0$. The derivatives of $\bar f \circ \varphi_x^{\tan}$ can be directly computed using a chain rule, as we will see shortly. However, another formulation can be obtained by examination of (2.4): substitute $\varphi_x^{\tan}(u)$ for $x + d$ in (2.4) particularized with $g = g_{\mathcal{U}}$, and use $\mathrm{P}_{\mathcal{U}}(\varphi_x^{\tan}(u) - x) = u$ to obtain

$$(f \circ \varphi_x^{\tan})(u) = f(x) + g_{\mathcal{U}}^\top u + \tfrac{1}{2} u^\top [\nabla^2 L_{\mathcal{U}}^{g_{\mathcal{U}}}(0)] u + O(\|u\|^3) \tag{3.5}$$

(similar to Theorem 4.1(v) in [19]). The expansion (3.5) is also straightforward from the expression $L_{\mathcal{U}}^{g_{\mathcal{U}}} = f \circ \varphi_x^{\tan}$ (since $g_{\mathcal{U}} = \nabla L_{\mathcal{U}}^{g_{\mathcal{U}}}(0)$). Then the $\mathcal{U}$-Hessian $\nabla^2 L_{\mathcal{U}}^{g_{\mathcal{U}}}(0)$ can also be obtained by computing derivatives of $\bar f \circ \varphi_x^{\tan}$, which leads to an interpretation of the quadratic term of this expansion. We have

$$\begin{aligned} \nabla(\bar f \circ \varphi_x^{\tan})(0)^\top \delta &= \nabla \bar f(x)^\top (I + \mathrm{D}v(0))\delta = \nabla \bar f(x)^\top \delta && \forall \delta \in \mathcal{U} \\ \delta_1^\top [\nabla^2(\bar f \circ \varphi_x^{\tan})(0)]\delta_2 &= \delta_1^\top \nabla^2 \bar f(x) \delta_2 + \nabla \bar f(x)^\top [\mathrm{H}v(0)](\delta_1, \delta_2) && \forall \delta_1, \delta_2 \in \mathcal{U}. \end{aligned}$$

We solve for $\mathrm{H}v(0)$ from (3.3). Introduce the matrices $M_1, \ldots, M_{n-p}$ (depending on x) such that

$$[\mathrm{H}\Phi(x)](\xi_1, \xi_2) = \begin{bmatrix} \xi_1^\top M_1 \xi_2 \\ \vdots \\ \xi_1^\top M_{n-p} \xi_2 \end{bmatrix} \quad \forall \xi_1, \xi_2 \in \mathbb{R}^n.$$

Let δ_1 and δ_2 be in $\mathcal{U}$. Then $[\mathrm{H}v(0)](\delta_1, \delta_2) \in \mathcal{V} = \mathrm{range}(\mathrm{D}\Phi(x)^*)$, hence

$$[\mathrm{H}v(0)](\delta_1, \delta_2) = -\mathrm{D}\Phi(x)^*(\mathrm{D}\Phi(x)\mathrm{D}\Phi(x)^*)^{-1}[\mathrm{H}\Phi(x)](\delta_1, \delta_2). \tag{3.6}$$

Finally, define

$$\lambda_{\mathrm{LS}} := -(\mathrm{D}\Phi(x)\mathrm{D}\Phi(x)^*)^{-1}\mathrm{D}\Phi(x)\nabla\bar{f}(x) \tag{3.7}$$

$$\bar{M} := \sum_{i=1}^{n-p}(\lambda_{\mathrm{LS}})_i M_i \tag{3.8}$$

so that

$$\delta_1^\top[\nabla^2(\bar{f}\circ\varphi_x^{\tan})(0)]\delta_2 = \delta_1^\top[\nabla^2\bar{f}(x) + \bar{M}]\delta_2.$$

We have derived the following second-order expansion of f on $\mathcal{M}$.

Theorem 3.4 (Tangential second-order expansion). *With $\bar{M}$ defined according to* (3.8)*, we have, for $u \in \mathcal{U}(x)$,*

$$(f\circ\varphi_x^{\tan})(u) = f(x) + \nabla\bar{f}(x)^\top u + \tfrac{1}{2}u^\top[\nabla^2\bar{f}(x) + \bar{M}]u + O(\|u\|^3). \tag{3.9}$$

Direct comparison of (3.5) and (3.9) gives

$$g_{\mathcal{U}} = \mathrm{P}_{\mathcal{U}}(\nabla\bar{f}(x)) \quad \text{and} \quad \nabla^2 L_{\mathcal{U}}^{g_{\mathcal{U}}}(0) = \mathrm{P}_{\mathcal{U}}[\nabla^2\bar{f}(x) + \bar{M}]\mathrm{P}_{\mathcal{U}}.$$

The first equality gives an intrinsic interpretation of the $\mathcal{U}$-gradient: $g_{\mathcal{U}}$ is the gradient of the restriction of f to $\mathcal{M}$. This implies in particular that

$$\nabla\bar{f}(x) \in \mathrm{aff}\ \partial f(x). \tag{3.10}$$

Given $\bar{f}$ and Φ, the second equality gives an explicit expression for the $\mathcal{U}$-Hessian $\nabla^2 L_{\mathcal{U}}^{g_{\mathcal{U}}}(0)$ when it exists. But note that Theorem 3.4 is not conditioned on $g_{\mathcal{U}} \in \mathrm{ri}\ \partial f(x)$. The $\mathcal{U}$-Newton direction based on the tangential parameterization is always well-defined, regardless of whether or not the $\mathcal{U}$-Hessian is.

Algorithm 3.5 ($\mathcal{U}$-Newton). *Let $x \in \mathcal{M}$ be given. Repeat:*

1. *Identify $\mathcal{U}(x)$ and a basis U of $\mathcal{U}(x)$. Compute $\nabla\bar{f}(x)$ and $\bar{M}$ from* (3.8)*.*
2. *Compute the Newton update:*

$$h_{\tan}(x) = -U\left(U^\top[\nabla^2\bar{f}(x) + \bar{M}]U\right)^{-1}U^\top\nabla\bar{f}(x), \tag{3.11}$$

$$N_{\tan}(x) = \varphi_x^{\tan}(h_{\tan}(x)) = x + h_{\tan}(x) + v(h_{\tan}(x)). \tag{3.12}$$

3. *Update $x \leftarrow N_{\tan}(x)$.*

Incidentally, notice that the choice of basis for $\mathcal{U}$ does not affect $h_{\tan}(x)$.

Remark 3.6. Observe that Algorithm 6 in [20] looks like a implementable version of this $\mathcal{U}$-Newton method. The $\mathcal{U}$-step uses an approximation of $\mathrm{P}_{\partial f(x)}(0)$ which is equal to $g_{\mathcal{U}}(x)$ if $x \in \mathcal{M}$ is close to a strong minimizer (see Corollary 2.14). The $\mathcal{V}$-step uses a bundle iteration to approximate the proximal point of $x + h_{\tan}(x)$, which if exact would put the next iterate back on the manifold. Neither step needs exact knowledge of the $\mathcal{U}$-gradient, the $\mathcal{U}$-Hessian, or $v(\cdot)$.

3.3. *The role of the $\mathcal{U}$-gradient as an interior subgradient*

We see from (3.12) that the $\mathcal{U}$-Newton step has two parts: a tangent step $h_{\tan} \in \mathcal{U}$, and a normal step $v(h_{\tan}) \in \mathcal{V}$ that puts $N_{\tan}(x)$ back on the manifold. Neither part is contingent on

$$g_{\mathcal{U}} \in \operatorname{ri} \partial f(x). \tag{3.13}$$

What this condition provides is the characterization of $v(u)$ in Theorem 2.5: for $u \in \mathcal{U}$ small enough,

$$v(u) = \operatorname*{argmin}_{v \in \mathcal{V}} f(x + u + v). \tag{3.14}$$

This is the $\mathcal{V}$-step proposed in the "conceptual superlinear scheme" of [11, Sect. 4.3], a variant of the $\mathcal{U}$-Newton method with $\mathcal{U}$ and $\mathcal{V}$ fixed at a strong minimizer $\bar{x}$ (so $g_{\mathcal{U}} = 0$). Nevertheless, (3.13) is not even necessary for the applicability of (3.14) in a superlinearly convergent algorithm. Mifflin and Sagastizábal [15] show that if $g_{\mathcal{U}} = 0 \notin \operatorname{ri} \partial f(\bar{x})$ so that $W^{g_{\mathcal{U}}}(u)$ (or even $W^{g_{\mathcal{U}}}(0)$) is not a singleton, taking a $\mathcal{V}$-step by selecting an arbitrary $v \in W^{g_{\mathcal{U}}}(u)$ still leads to superlinear convergence, as long as a linear growth condition on $u \mapsto W^{g_{\mathcal{U}}}(u)$ holds (satisfied in our situation thanks to Corollary 2.3).

It is worth noting, however, that the proximal-point approximation to a $\mathcal{U}$-Newton scheme in [20] requires the existence of a strong minimizer, and in particular $0 \in \operatorname{ri} \partial f(\bar{x})$.

An alternative to (3.14) exists for implementing a $\mathcal{V}$-step. Oustry [21] uses a projection onto the manifold in the "$\mathcal{U}$-Newton" algorithm for minimizing the maximum eigenvalue function (see [21, Alg. 6.4]). This is not a $\mathcal{V}$-step in our sense, since it is not perpendicular to $\mathcal{U}$ but rather to the manifold at the projected point. Nevertheless, we will show, in Theorem 4.9, that the steps are the same to second order. At this point, we only wish to note that the subgradient $g = \mathrm{P}_{\partial f(x)}(0)$ selected for defining the $\mathcal{U}$-Hessian in [21] is actually $g_{\mathcal{U}}$ for x close enough to $\bar{x}$. In general $g_{\mathcal{U}} \neq g$, but the convergence analysis in [21] assumes that $0 \in \operatorname{ri} \partial f(\bar{x})$, which by Corollary 2.14 guarantees that $g_{\mathcal{U}} = g$. The $\mathcal{U}$-Hessian defined by g thus correctly reflects the curvature of f on $\mathcal{M}$ near the minimizer, and leads to the quadratic convergence of [21, Alg. 6.4].

4. The Riemannian Newton Method

The tangential parameterization of the $\mathcal{U}$-Newton method is not the only parameterization of $\mathcal{M}$. Others lead to different Newton steps. However, in this section we demonstrate the intrinsic nature of the $\mathcal{U}$-Newton *direction* by comparing it to a sequential Newton method defined using Riemannian geometry. This connection also provides a proof of local quadratic convergence of $\mathcal{U}$-Newton.

4.1. *Geodesics*

Since $\mathcal{M}$ is a differentiable manifold, it may be endowed with a Riemannian metric to make it a Riemannian manifold. The Riemannian Newton method uses geodesics to parameterize the manifold $\mathcal{M}$. Roughly speaking, geodesics are length-minimizing curves among those traced with constant speed (see among others [10, Ch. 4] or [5, Ch. 3]).

Let $y(t)$ be a smooth path in $\mathcal{M}$ with real parameter t, and suppose it is traced with constant speed ($\|\dot{y}(t)\|$ is constant for all t). Because $\mathcal{M}$ is embedded in $\mathbb{R}^n$, such a path $y(t)$ is a geodesic if and only if its acceleration at any point t is normal to the manifold at $y(t)$. Hence we differentiate the local equations for the manifold $\Phi(y) = 0$ twice to obtain local equations for a geodesic. The argument (t) will sometimes be suppressed for brevity.

$$\frac{d}{dt}\Phi(y(t)) = \mathrm{D}\Phi(y)\dot{y} = 0 \tag{4.1}$$

$$\frac{d^2}{dt^2}\Phi(y(t)) = \mathrm{D}\Phi(y)\ddot{y} + [\mathrm{H}\Phi(y)](\dot{y}, \dot{y}) = 0. \tag{4.2}$$

For $y(t)$ to be a geodesic, $\ddot{y}$ must be a normal vector, which according to Lemma 2.1 means

$$\exists\, \theta \in \mathbb{R}^{n-p} \text{ such that } \quad \ddot{y} = \mathrm{D}\Phi(y)^*\theta. \tag{4.3}$$

Combining (4.2) and (4.3) and solving for $\ddot{y}$, we get the differential equation for a geodesic (see the correspondence with (3.6)):

$$\ddot{y} = -\mathrm{D}\Phi(y)^* \left(\mathrm{D}\Phi(y)\mathrm{D}\Phi(y)^*\right)^{-1} [\mathrm{H}\Phi(y)](\dot{y}, \dot{y}). \tag{4.4}$$

Notice that the inverse of $\mathrm{D}\Phi(y)\mathrm{D}\Phi(y)^*$ exists for y close enough to x by surjectivity of $\mathrm{D}\Phi(x)$ and smoothness of Φ. The solutions of interest are those with $y(0) = x$, and $\dot{y}(0)$ in $\mathcal{U}$ because of (4.1). Existence and uniqueness of maximal solutions are assured (see [10, Th. 4.10]). Therefore, the free parameter $u \in \mathcal{U}$ locally determines the geodesic through initial conditions $y(0) = x$ and $\dot{y}(0) = u$. We adopt the classical notation $\gamma(t, x, u)$ for this geodesic. The uniqueness also yields the rescaling property ([5, Ch. 3, Lemma 2.6]): for $a \in \mathbb{R}$,

$$\gamma(t, x, au) = \gamma(at, x, u) \tag{4.5}$$

whenever either side is defined. The function γ can be used to define the *exponential parameterization* of $\mathcal{M}$ at x.

Lemma 4.1 (Exponential parameterization). *The function $\varphi_x^{\exp}$ defined by*

$$\varphi_x^{\exp}(u) := \gamma(1, x, u) \qquad \textit{for all } u \in \mathcal{U}(x) \textit{ small enough}$$

is a smooth local parameterization of $\mathcal{M}$ around x. Moreover $\{\varphi_x^{\exp}\}_x$ is a smooth parameterization family.

Proof. The function $\varphi_x^{\exp}$ is exactly the *exponential map* (usually denoted $\exp_x$), a standard parameterization in Riemannian geometry. See [5, Ch. 3, Prop. 2.7] for the smoothness of $(x, u) \mapsto \gamma(1, x, u)$, and [5, Ch. 3, Prop. 2.9] for the parameterization $\exp_x$. □

4.2. Extrinsic and intrinsic Riemannian Newton

An extrinsic formulation of the Riemannian Newton method comes from Algorithm 1.8 using $\varphi_x^{\exp}$ as the parameterization.

Algorithm 4.2 (Riemannian Newton). *Given a point $x \in \mathcal{M}$, repeat the update $x \leftarrow N_{\exp}(x)$ where*

$$\begin{aligned} h_{\exp}(x) &= -\left[\nabla^2(\bar{f} \circ \varphi_x^{\exp})(0)\right]^{-1} \nabla(\bar{f} \circ \varphi_x^{\exp})(0) \qquad (4.6)\\ N_{\exp}(x) &= \varphi_x^{\exp}(h_{\exp}(x)). \end{aligned}$$

A Newton method may also be formulated using *covariant derivatives*, which are intrinsic geometric objects that express the derivatives of a function or vector field on a differentiable manifold (see e.g. [5, 10]). Using

$$\nabla_{\mathcal{M}} f(x) \in \mathrm{T}_{\mathcal{M}}(x) \qquad \text{and} \qquad \nabla^2_{\mathcal{M}} f(x) : \mathrm{T}_{\mathcal{M}}(x) \to \mathrm{T}_{\mathcal{M}}(x)$$

to denote respectively the covariant derivative and Hessian of f on $\mathcal{M}$ at x, the intrinsic Riemannian Newton method is the iteration $x \leftarrow N_{\mathrm{R}}(x)$ where

$$h_{\mathrm{R}}(x) = -[\nabla^2_{\mathcal{M}} f(x)]^{-1} \nabla_{\mathcal{M}} f(x) \qquad (4.7)$$

$$N_{\mathrm{R}}(x) = \varphi_x^{\exp}(h_{\mathrm{R}}(x)). \qquad (4.8)$$

This iteration has appeared in particular in [7, 26, 27, 6, 3].

The connection between the intrinsic and extrinsic Riemannian Newton methods can be established via the following Taylor formula (see Remark 3.2 in [26] for instance):

$$f(\varphi_x^{\exp}(tu)) = f(x) + t\nabla_{\mathcal{M}} f(x)^\top u + \tfrac{1}{2} t^2 u^\top [\nabla^2_{\mathcal{M}} f(x)] u + O(t^3). \qquad (4.9)$$

Thus, we have

$$\nabla_{\mathcal{M}} f(x) = \nabla(\bar{f} \circ \varphi_x^{\exp})(0) \qquad \text{and} \qquad \nabla^2_{\mathcal{M}} f(x) = \nabla^2(\bar{f} \circ \varphi_x^{\exp})(0), \quad (4.10)$$

so $h_{\exp} = h_{\mathrm{R}}$ and the two formulations of Riemannian Newton are identical.

Local quadratic convergence of Riemannian Newton was proved in particular in [7] and [26] using geometric arguments. We present a simple proof that will be useful as a model for the other sequential Newton methods in this paper.

Lemma 4.3 (Quadratic convergence of sequential Newton). *Let $\bar{x} \in \mathcal{M}$ be such that $\nabla_{\mathcal{M}} f(\bar{x}) = 0$ and $\nabla^2_{\mathcal{M}} f(\bar{x})$ is nonsingular. Suppose $\{\varphi_x\}_x$ is a smooth parameterization family of $\mathcal{M}$ with $\mathrm{D}\varphi_x(0) = I$ for all $x \in \mathcal{M}$. Then for x close enough to $\bar{x}$, the Newton step $h_{\mathrm{R}}(x)$ is well-defined and the Newton update given by*

$$N(x) = \varphi_x(h_{\mathrm{R}}(x)) \qquad (4.11)$$

is quadratically closer to $\bar{x}$ than x is.

Proof. First note that the smoothness of f and the invertibility of $\nabla^2_{\mathcal{M}} f(\bar{x})$ yield that $\nabla^2_{\mathcal{M}} f(x)$ is nonsingular (and then $h_{\mathrm{R}}(x)$ is well-defined) in some neighborhood Ω of $\bar{x}$ in $\mathcal{M}$. Observe also that $h_{\mathrm{R}}(\bar{x}) = 0$ and $N(\bar{x}) = \bar{x}$; that is, $\bar{x}$ is a fixed point of N.

Since N is smooth, its first-order expansion at $\bar{x}$ is

$$N(x) = N(\bar{x}) + \mathrm{D}N(\bar{x})(x - \bar{x}) + O(\|x - \bar{x}\|^2).$$

We proceed to show $\mathrm{D}N(\bar{x}) = 0$, since it then follows that

$$N(x) - \bar{x} = O(\|x - \bar{x}\|^2).$$

Define $A(x) := [\nabla^2_{\mathcal{M}} f(x)]^{-1}$ for $x \in \Omega$, so $h_{\mathrm{R}}(x) = -A(x)\nabla_{\mathcal{M}} f(x)$. Compute the covariant directional derivative of h_{R} at $\bar{x}$ in direction $u \in \mathrm{T}_{\mathcal{M}}(\bar{x})$:

$$h'_{\mathrm{R}}(\bar{x}; u) = -A'(\bar{x}; u)\nabla_{\mathcal{M}} f(\bar{x}) - A(\bar{x})[\nabla^2_{\mathcal{M}} f(\bar{x})]u = -u,$$

which shows $\mathrm{D}h_{\mathrm{R}}(\bar{x}) = -I$. Now define $\bar{h}(x) := (x,\ h_{\mathrm{R}}(x)) \in \mathrm{T}\mathcal{M}$, and $\varphi(x, u) := \varphi_x(u)$ for $(x, u) \in \mathrm{T}\mathcal{M}$. Then $N = \varphi \circ \bar{h}$. Identify $\mathrm{T}_{\mathrm{T}\mathcal{M}}(x) = \mathrm{T}_{\mathcal{M}}(x) \times \mathrm{T}_{\mathcal{M}}(x)$, and observe that $\mathrm{D}\varphi(x, 0) = [I \;\; I]$ for any $x \in \mathcal{M}$. Hence

$$\mathrm{D}N(\bar{x}) \;=\; \mathrm{D}\varphi(\bar{h}(\bar{x}))\,\mathrm{D}\bar{h}(\bar{x}) \;=\; \begin{bmatrix} I & I \end{bmatrix}\begin{bmatrix} I \\ -I \end{bmatrix} = 0,$$

which completes the proof. □

Theorem 4.4 (Quadratic convergence of Riemannian Newton). *Let $\bar{x} \in \mathcal{M}$ be such that $\nabla(\bar{f} \circ \varphi^{\exp}_{\bar{x}})(0) = 0$ and $\nabla^2(\bar{f} \circ \varphi^{\exp}_{\bar{x}})(0)$ is nonsingular. Then for x close enough to $\bar{x}$, the Riemannian Newton step $h_{\exp}(x)$ in Algorithm 4.2 is well-defined, and the Riemannian Newton update $N_{\exp}(x)$ is quadratically closer to $\bar{x}$ than x is.*

Proof. Given $u \in \mathrm{T}_{\mathcal{M}}(x)$, the directional derivative of $\varphi^{\exp}_x$ at 0 is

$$\varphi^{\exp\prime}_x(0; u) = \frac{d}{dt}\gamma(1, x, tu)\Big|_{t=0} = \frac{d}{dt}\gamma(t, x, u)\Big|_{t=0} = u,$$

using (4.5) with (t, u) there set to $(1, t)$. Thus $\mathrm{D}\varphi^{\exp}_x(0) = I$. Using (4.10) and $h_{\exp} = h_{\mathrm{R}}$, apply Lemma 4.3 with $\varphi_x = \varphi^{\exp}_x$ to complete the proof. □

4.3. Connection to $\mathcal{U}$-Newton

The tangentially parameterized paths of the $\mathcal{U}$-Newton method are not geodesics in general. However, they are close to them.

Theorem 4.5 (Tangential parameterization and geodesics). *Tangentially parameterized paths agree with geodesics up to second-order: for $x \in \mathcal{M}$ and $u \in \mathcal{U}(x)$,*

$$\gamma(t, x, u) = x + tu + v(x, tu) + o(t^2). \tag{4.12}$$

Proof. Let $x \in \mathcal{M}$ and $u \in \mathcal{U}$. For t small enough, let $y(t) := \gamma(t, x, u)$ be the geodesic given by u, and let $z(t) := x + tu + v(tu)$ be the tangentially parameterized curve given by u. (Since x is fixed, the dependence of v on x is dropped for notational convenience.) We have $z(0) = x$, and $\dot{z}(t) = u + \mathrm{D}v(tu)u$ so $\dot{z}(0) = u$ (by Corollary 2.3). Thus $y(t)$ and $z(t)$ agree to first order. Now recall that being a geodesic requires $\ddot{y}(0) \in \mathcal{V}$. This implies that

$$\mathrm{P}_{\mathcal{U}}(y(t) - x) = tu + o(t^2).$$

Therefore Corollary 2.3 yields that

$$\mathrm{P}_{\mathcal{V}}(y(t) - x) = v(tu + o(t^2)).$$

Since v is smooth, it is locally Lipschitz in particular and then

$$\mathrm{P}_{\mathcal{V}}(y(t) - x) = v(tu) + o(t^2).$$

Finally we get

$$y(t) \;=\; x + \mathrm{P}_{\mathcal{U}}(y(t) - x) + \mathrm{P}_{\mathcal{V}}(y(t) - x) \;=\; x + tu + v(tu) + o(t^2),$$

which ends the proof. □

The first corollary is about fast tracks: Theorem 2.9 states that a convex function admitting a fast track is actually partly smooth relative to this fast track. Thus parameterized paths $\bar{x} + tu + w(tu)$ on a fast track are geodesics up to second order. A second corollary is the following.

Corollary 4.6. *The exponential and tangential parameterizations agree to second order:*

$$\varphi_x^{\exp}(u) = \varphi_x^{\tan}(u) + o(\|u\|^2).$$

Proof. For $u \neq 0$, the rescaling property (4.5) yields $\gamma(1, x, u) = \gamma(\|u\|, x, \frac{u}{\|u\|})$ so we can rewrite (4.12) as $\gamma(1, x, u) = x + u + v(u) + o(\|u\|^2)$. □

A consequence of this corollary and equations (3.5) and (4.9) is that the first and second covariant derivatives of f on $\mathcal{M}$ at x may be computed from $\mathcal{U}$-objects:

$$\nabla_{\mathcal{M}} f(x) = g_{\mathcal{U}} \qquad \text{and} \qquad \nabla^2_{\mathcal{M}} f(x) = \nabla^2 L_{\mathcal{U}}^{g_{\mathcal{U}}}(0). \tag{4.13}$$

Furthermore, Corollary 4.6 implies that $\mathcal{U}$-Newton gives the same Newton direction as Riemannian Newton, and local quadratic convergence is preserved.

Theorem 4.7 (Quadratic convergence of $\mathcal{U}$-Newton). *Let $\bar{x} \in \mathcal{M}$ be such that $g_{\mathcal{U}} = 0$ and $\nabla^2 L_{\mathcal{U}}^{g_{\mathcal{U}}}(0)$ is nonsingular. Then for x close enough to $\bar{x}$, the $\mathcal{U}$-Newton step $h_{\tan}(x)$ given by Algorithm 3.5 is well-defined, and the $\mathcal{U}$-Newton update $N_{\tan}(x)$ is quadratically closer to $\bar{x}$ than x is.*

Proof. Observe Theorem 2.2 and (3.4) imply $\mathrm{D}\varphi_x^{\tan}(0) = I$. Moreover Corollary 4.6 implies that $h_{\tan} = h_{\exp} = h_{\mathrm{R}}$, since the Newton direction depends only on first and second derivatives. Finally (4.13) ensures that we can apply Lemma 4.3 with $\varphi_x = \varphi_x^{\tan}$ to complete the proof. □

4.4. Connection to projected $\mathcal{U}$-Newton

We refer to the "$\mathcal{U}$-Newton method" in [21] as the *projected $\mathcal{U}$-Newton method* because it projects onto $\mathcal{M}$ in place of the $\mathcal{V}$-step from the tangential parameterization. This method implicitly uses a different parameterization of the manifold, what might be called a *projection parameterization*: for $x \in \mathcal{M}$ and $u \in \mathcal{U}(x)$

$$\varphi_x^{\text{proj}}(u) := \mathrm{P}_{\mathcal{M}}(x+u). \tag{4.14}$$

The projection is well-defined for u small enough. Let us prove that it is a parameterization of $\mathcal{M}$ and that it reproduces the exponential parameterization to second order.

Lemma 4.8 (Projection parameterization). *The function φ_x^{proj} is a smooth local parameterization of $\mathcal{M}$ around $x \in \mathcal{M}$ and the family $\{\varphi_x^{\text{proj}}\}_x$ is a smooth parameterization family.*

Proof. Recall that $\mathrm{N}\mathcal{M} = \{(x,v) \in \mathbb{R}^n \times \mathbb{R}^n \mid x \in \mathcal{M},\ v \in \mathrm{N}_{\mathcal{M}}(x)\}$ is a smooth manifold of dimension n. It is easily shown using the equations for $\mathrm{N}\mathcal{M}$ that

$$\mathrm{T}_{\mathrm{N}\mathcal{M}}(x,0) = \mathrm{T}_{\mathcal{M}}(x) \times \mathrm{N}_{\mathcal{M}}(x).$$

Consider the smooth function

$$F\colon \begin{cases} \mathrm{N}\mathcal{M} \longrightarrow \mathbb{R}^n \\ (x,v) \longmapsto x+v\,. \end{cases}$$

Its derivative $\mathrm{D}F(x,0) : \mathrm{T}_{\mathrm{N}\mathcal{M}}(x,0) \to \mathbb{R}^n$ is given by

$$\mathrm{D}F(x,0)\begin{bmatrix} u \\ v \end{bmatrix} = u + v$$

so it is obviously invertible. Thus the local inverse theorem (for manifolds) yields that F is a local diffeomorphism from a neighborhood of $(x,0)$ in $\mathrm{N}\mathcal{M}$ into its image. Introducing the projection

$$\pi_1\colon \begin{cases} \mathrm{N}\mathcal{M} \longrightarrow \mathcal{M} \\ (x,v) \longmapsto x\,, \end{cases}$$

the function $\psi = \pi_1 \circ F^{-1}$ defined from a neighborhood of x in $\mathbb{R}^n$ to a neighborhood of x in $\mathcal{M}$ is also smooth. Now observe that φ_x^{proj} is the restriction of ψ to $x + \mathrm{T}_{\mathcal{M}}(x)$. Thus we can write

$$\varphi_x^{\text{proj}}(u) = \pi_1(F^{-1}(x+u)). \tag{4.15}$$

First conclusions are that φ_x^{proj} is smooth and that $(x,u) \mapsto \varphi_x^{\text{proj}}(u)$ is smooth too. Taking the derivative of (4.15), we get for all $u \in \mathcal{U}(x)$,

$$\mathrm{D}\varphi_x^{\text{proj}}(0)u = \pi_1(u,0) = u. \tag{4.16}$$

Then the local inverse theorem yields that φ_x^{proj} is a smooth diffeomorphism, and thus φ_x^{proj} is a local parameterization of $\mathcal{M}$ around x. □

Theorem 4.9. *The projection parameterization agrees with the exponential parameterization to second order: for $u \in \mathcal{U}(x)$ small enough*

$$\varphi_x^{\exp}(u) = \varphi_x^{\text{proj}}(u) + o(\|u\|^2).$$

Proof. In this proof we denote, for $x \in \mathcal{M}$ and $u \in \mathcal{U}$, by $y(t) = \gamma(t, x, u)$ the geodesic satisfying $y(0) = x$ and $\dot{y}(0) = u$, and by $H_x(u, u)$ the second fundamental form (see [10, Ch. 8] for example). Fix $x \in \mathcal{M}$ and $u \in \mathcal{U}$ and set

$$\theta(t) := y(t) - \tfrac{1}{2}t^2\ H_{y(t)}(\dot{y}(t), \dot{y}(t)).$$

Observe that

$$\theta(0) = x, \quad \dot{\theta}(0) = u, \quad \text{and} \quad \ddot{\theta}(0) = \ddot{y}(0) - H_x(u, u).$$

Since $y(t)$ is a geodesic, the Gauss Formula [10, Lemma 8.5] enables to write $\ddot{y}(0) = 0 + H_x(u, u)$, and then $\ddot{\theta}(0) = 0$. Hence

$$\theta(t) = x + tu + o(t^2).$$

Projecting into $\mathcal{M}$, we get

$$\mathrm{P}_{\mathcal{M}}(\theta(t)) = \mathrm{P}_{\mathcal{M}}(x + tu) + o(t^2).$$

Since $y(t) \in \mathcal{M}$ and $H_{y(t)}(\dot{y}(t), \dot{y}(t)) \in \mathrm{N}\mathcal{M}(y(t))$, we can write

$$\mathrm{P}_{\mathcal{M}}(\theta(t)) = \mathrm{P}_{\mathcal{M}}\big(y(t) - \frac{t^2}{2}H_{y(t)}(\dot{y}(t), \dot{y}(t))\big) = \mathrm{P}_{\mathcal{M}}(y(t)) \ = \ y(t).$$

Finally, we thus have

$$y(t) = \varphi_x^{\text{proj}}(tu) + o(t^2).$$

Using rescaling lemma (as in Corollary 4.6), we obtain that the projection parameterization agrees with the exponential parameterization to second order. □

The following summarizes the relations between the three parameterizations considered in the paper : for $u \in \mathcal{U}(x)$,

$$\gamma(t, x, u) = \varphi_x^{\exp}(tu) = \varphi_x^{\tan}(tu) + o(t^2) = \varphi_x^{\text{proj}}(tu) + o(t^2).$$

Theorem 4.10 (Quadratic convergence of projected $\mathcal{U}$-Newton). *Let $\bar{x} \in \mathcal{M}$ such that $\nabla(\bar{f} \circ \varphi_{\bar{x}}^{\text{proj}})(0) = 0$ and $\nabla^2(\bar{f} \circ \varphi_{\bar{x}}^{\text{proj}})(0)$ is nonsingular. Then for x close enough to $\bar{x}$, the projected $\mathcal{U}$-Newton step*

$$h_{\text{proj}}(x) = -\left[\nabla^2(\bar{f} \circ \varphi_x^{\text{proj}})(0)\right]^{-1} \nabla(\bar{f} \circ \varphi_x^{\text{proj}})(0)$$

is well-defined, and the projected $\mathcal{U}$-Newton update

$$N_{\text{proj}}(x) = \varphi_x^{\text{proj}}(h_{\text{proj}}(x))$$

is quadratically closer to $\bar{x}$ than x is.

Proof. Equation (4.16) implies $\mathrm{D}\varphi_x^{\text{proj}}(0) = I$. In addition, Theorem 4.9 implies that $h_{\text{proj}} = h_{\exp} = h_{\mathrm{R}}$, since the Newton direction depends only on first and second derivatives. Then Lemma 4.3 with $\varphi_x = \varphi_x^{\text{proj}}$ completes the proof. □

5. Sequential Quadratic Programming

Recall the framework set in the introduction: represent the manifold $\mathcal{M}$ around x by a local equation $\{y \mid \Phi(y) = 0\}$, and replace f by a smooth function $\bar{f}$ that coincides with f on $\mathcal{M}$. Now we solve (1.2), which is locally equivalent to the original problem of minimizing f on $\mathcal{M}$. The Lagrangian for this problem is

$$L(y, \lambda) := \bar{f}(y) + \lambda^{\top}\Phi(y),$$

and the first-order optimality conditions are

$$\nabla_y L(y, \lambda) = \nabla \bar{f}(y) + \mathrm{D}\Phi^*(y)\lambda = 0 \tag{5.1}$$

$$\nabla_\lambda L(y, \lambda) = \Phi(y) = 0. \tag{5.2}$$

Each iteration of the SQP method solves a linearization of (5.1)–(5.2), linearized at $y = x$ and some choice of λ intended to approximate the optimal Lagrange multipliers (see [1] for instance). With a change of variables, this can be shown to be the same as solving the quadratic program

$$\begin{array}{ll} \min\limits_{d} & \nabla \bar{f}(x)^{\top} d + \frac{1}{2} d^{\top}[\nabla_y^2 L(x, \lambda)]d \\ \text{s.t.} & \Phi(x) + \mathrm{D}\Phi(x)d = 0 \end{array} \tag{5.3}$$

as long as $d^{\top}[\nabla_y^2 L(x, \lambda)]d > 0$ for feasible directions d.

Of course, this Hessian depends on the choice of λ, which is reminiscent of the dependence of the $\mathcal{U}$-Hessian on g. In fact, for any $g_0 \in \operatorname{aff} \partial f(x)$, there is a one-to-one correspondence between $\lambda \in \mathbb{R}^{n-p}$ and $g \in \operatorname{aff} \partial f(x)$ through the relation

$$g = g_0 + \mathrm{D}\Phi(x)^*\lambda, \tag{5.4}$$

although g and g_0 may not be subgradients. Two common choices for λ are

(i) the optimal Lagrange multipliers from the quadratic program of the previous iteration,
(ii) the multipliers that solve the least-squares problem

$$\min_{\lambda} \left\| \nabla \bar{f}(x) + \mathrm{D}\Phi(x)^*\lambda \right\|^2, \tag{5.5}$$

approximately solving the optimality condition (5.1) at $y = x$.

The first choice results in a Newton method for solving the optimality conditions (5.1)–(5.2), producing quadratic convergence of the (x, λ) iterates [1, Th. 13.2], so it is well-motivated analytically. The second choice has a more geometric motivation, as explained in the following theorem (which uses the notation of Section 3.2).

Theorem 5.1 (SQP and $\mathcal{U}$-Newton). *Let $x \in \mathcal{M}$ and $\lambda = \lambda(x)$ solve* (5.5). *Hence*

$$g_{\mathcal{U}}(x) = \nabla \bar{f}(x) + \mathrm{D}\Phi(x)^*\lambda(x). \tag{5.6}$$

Moreover, the next iterate computed by SQP is the $\mathcal{U}$-step of the $\mathcal{U}$-Newton algorithm: if $\bar{M} + \nabla^2 \bar{f}(x)$ is positive definite on $\mathcal{U}(x)$, the solution to (5.3) *exists and is $d = h_{\tan}(x) = h_{\exp}(x) = h_{\mathrm{R}}(x) = h_{\mathrm{proj}}(x)$.*

Proof. By (3.10), set $g_0 = \nabla \bar{f}(x)$ in (5.4). Then solving (5.5) corresponds to projecting 0 on aff $\partial f(x)$. By Lemma 2.4, this yields (5.6). As $x \in \mathcal{M}$, the constraint of (5.3) can be written $\mathrm{D}\Phi(x)d = 0$, which means $d \in \mathcal{U}$ (see Lemma 2.1). Now observe that λ_{LS} defined in (3.7) solves (5.5) (these are the least-squares multipliers, as suggested by the notation). Hence $\nabla^2_y L(x, \lambda_{\mathrm{LS}}) = \nabla^2 \bar{f}(x) + \bar{M}$, and thus (5.3) is equivalent to

$$\min_{u \in \mathcal{U}} \; \nabla \bar{f}(x)^\top u + \tfrac{1}{2} u^\top [\nabla^2 \bar{f}(x) + \bar{M}] u$$

when $\lambda = \lambda_{\mathrm{LS}}$. The solution of this quadratic program is $u = h_{\tan}(x)$ defined in (3.11). We can conclude using the definitions of $h_{\exp}$, h_{R} and h_{proj}, and Corollary 4.6 and Theorem 4.9. □

This theorem provides a geometric interpretation for the direction computed by SQP when least-squares multipliers are used: it is the Newton direction for the function f constrained on the manifold $\mathcal{M}$. There is a key difference with SQP, though. The constrained Newton methods discussed here are valid only *on* the manifold, whereas SQP is valid over the whole space $\mathbb{R}^n$. In SQP, x is updated to $x + d$ with d solving (5.3), without an explicit attempt to restore x to the manifold. SQP is intended to achieve both feasibility and optimality asymptotically.

However, this difference is not as meaningful as it might seem. Like any Newton method, SQP should be "globalized" with the help of a line-search or trust-region technique, both using a merit function q. As a rule, q is nonsmooth; typically one takes $q(y) = f(y) + \pi \|\Phi(y)\|$ with π large enough. It may happen that, no matter how close x is to a solution x^* (a point minimizing q), the Newton iterate $x + d$ (quadratically closer to x^*) may have $q(x + d) > q(x)$, and therefore be rejected by the line search. This so-called "Maratos effect" can cause a loss of superlinear convergence, and it has more chance to occur when $x \in \mathcal{M}$; see an illustration in [2].

The remedy proposed by Maratos himself [13] is to make a move toward $\mathcal{M}$ using a second-order model of the constraints. A more common and simpler remedy, apparently first proposed by Mayne in [14], uses a first-order model of the constraints. One takes an additional step $v \in \mathcal{V}(x)$ computed according to

$$v = -\mathrm{D}\Phi(x)^* \left(\mathrm{D}\Phi(x)\mathrm{D}\Phi(x)^*\right)^{-1} \Phi(x + d) \tag{5.7}$$

and updates x to $x + d + v$. We have seen from Theorem 5.1 that d is the $\mathcal{U}$-portion of the $\mathcal{U}$-Newton step; the correction (5.7) is itself a sort of Newton approximation to the $\mathcal{V}$ portion of the $\mathcal{U}$-Newton step. Thus, even though SQP constructs its iterates in the whole space, convergence rate may be improved by restricting it to a method staying near $\mathcal{M}$, approximating the $\mathcal{U}$-Newton method.

The globally convergent algorithm of [20], being a different approximation of $\mathcal{U}$-Newton (see Remark 3.6), presents another remedy to the Maratos effect. The iterates stay near $\mathcal{M}$ (the fast track) via adequate approximations of proximal steps, using a bundling mechanism that needs neither $\bar{f}$ nor Φ. As a referee has pointed out, the good numerical results of this algorithm may result from the bundle stopping test generating good least-squares multipliers.

Acknowledgements. The first author wishes to acknowledge the support of Claude Lemaréchal at INRIA Rhône-Alpes and Thomas Bewley at UCSD during the conception and completion of this work. Both authors would like to thank Claude Lemaréchal for fruitful discussions, and the anonymous referees for their careful reading and detailed suggestions.

References

1. Bonnans, J., Gilbert, J., Lemaréchal, C., Sagastizábal, C.: Numerical Optimization. Universitext. Springer-Verlag, 2003
2. Chamberlain, R., Lemaréchal, C., Pedersen, H., Powell, M.: The watchdog technique for forcing convergence in algorithms for constrained optimization. Math. Program. Study Series **16**, 1–17 (1982)
3. Dedieu, J.-P., Priouret, P., Malajovich, G.: Newton's method on Riemannian manifolds: covariant alpha-theory. IMA J. Numer. Anal. **23** (3), 395–419 (2003)
4. Dennis, J.E., Moré, J.J.: Quasi-Newton methods, motivation and theory. SIAM Review **19**(1), 46–89 (1977)
5. do Carmo, M.P.: Riemannian Geometry. Mathematics: Theory and Applications. Birkhäuser, 1992
6. Edelman, A., Arias, T., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20** (2), 303–353 (1999)
7. Gabay, D.: Minimizing a differentiable function over a differentiable manifold. J. Optimization Theory Appl. **37**(2), 177–219 June 1982
8. Hare, W.: Recent functions and sets of smooth substructure: Relationships and examples. To appear in J. Comput. Optim. Appl. available at `http://www.cecm.sfu.ca/~whare/Subsmooth-relations.ps`, Mar. 17, 2005
9. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms. Number 305–306 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1993
10. Lee, J.M.: Riemannian Manifolds: An Introduction to Curvature. Number 176 in Graduate Texts in Mathematics. Springer-Verlag, New York, 1997
11. Lemaréchal, C., Oustry, F., Sagastizábal, C.: The $\mathcal{U}$-Lagrangian of a convex function. Trans. AMS **352**(2), 711–729 (1999)
12. Lewis, A.S.: Active sets, nonsmoothness and sensitivity. SIAM J. Optim. **13**, 702–725 (2003)
13. Maratos, N.: Exact penalty function algorithms for finite dimensional and control optimization problems. PhD thesis, Imperial College, London, 1978
14. Mayne, D.Q.: On the use of exact penalty functions to determine step length in optimization algorithms. In: Lecture Notes in Mathematics, volume **773**, Springer Verlag, 1980 pp 98–109
15. Mifflin, R., Sagastizábal, C.: $\mathcal{VU}$-decomposition derivatives for convex max-functions. In: M. Théra and R. Tichatschke, editors, Ill-Posed Variational Problems and Regularization Techniques, Springer-Verlag, Berlin, 1999 pp 167–186
16. Mifflin, R., Sagastizábal C.: On $\mathcal{VU}$-theory for functions with primal-dual gradient structure. SIAM J. Optim. **11** (2), 547–571 (2000)
17. Mifflin, R., Sagastizábal, C.: Proximal points are on the fast track. J. Convex Anal. **9** (2), 563–579 (2002)
18. Mifflin, R., Sagastizábal, C.: Primal-dual gradient structured functions: second-order results; links to epi-derivatives and partly smooth functions. SIAM J. Optim. **13** (4), 1174–1194 (2003)
19. Mifflin, R., Sagastizábal, C.: $\mathcal{VU}$-smoothness and proximal point results for some nonconvex functions. Optim. Methods Softw. **19** (5), 463–478 (2004)
20. Mifflin, R., Sagastizábal, C.: A $\mathcal{VU}$-algorithm for convex minimization. Preprint, available at `http://www.sci.wsu.edu/math/faculty/mifflin/vualgo.ps`, revised Feb. 14, 2005
21. Oustry, F.: The $\mathcal{U}$-Lagrangian of the maximum eigenvalue function. SIAM J. Optim. **9** (2), 526–549 (1999)
22. Overton, M.L.: Large-scale optimization of eigenvalues. SIAM J. Optim. **2**, 88–120 (1992)
23. Overton, M.L., Ye, X.: Towards second-order methods for structured nonsmooth optimization. In: S. Gomez and J. P. Hennart, editors, Advances in Optimization and Numerical Analysis, Kluwer, 1994 pp 97–110
24. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Number 317 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1998
25. Shapiro, A., Fan, M.K.H.: On eigenvalue optimization. SIAM J. Optim. **5** (3), 552–569 (1995)
26. Smith, S.T.: Optimization techniques on Riemannian manifolds. Fields Inst. Comm. **3**, 113–136 (1994)
27. Udrişte, C.: Convex functions and optimization methods on Riemannian manifolds. Number 297 in Mathematics and its Applications. Kluwer, 1994