manuscript No.
(will be inserted by the editor)

Optimality conditions for the nonlinear programming problems on Riemannian manifolds

Wei Hong Yang · **Lei-Hong Zhang** · **Ruyi Song**

Abstract In recent years, many traditional optimization methods have been successfully generalized to minimize objective functions on manifolds. In this paper, we first extend the general traditional constrained optimization problem to a nonlinear programming problem built upon a general Riemannian manifold $\mathscr{M}$, and discuss the first-order and the second-order optimality conditions. By exploiting the differential geometry structure of the underlying manifold $\mathscr{M}$, we show that, in the language of differential geometry, the first-order and the second-order optimality conditions of the nonlinear programming problem on $\mathscr{M}$ coincide with the traditional optimality conditions. When the objective function is nonsmooth Lipschitz continuous, we extend the Clarke generalized gradient, tangent and normal cone, and establish the first-order optimality conditions. For the case when $\mathscr{M}$ is an embedded submanifold of $\mathbb{R}^m$, formed by a set of equality constraints, we show that the optimality conditions can be derived directly from the traditional results on $\mathbb{R}^m$.

Keywords Nonlinear programming · Optimality condition · Riemannian manifold · Generalized gradient · Hessian

Mathematics Subject Classification (2000) 90C30 · 90C46 · 65K05 · 58C05 · 49K27 · 49M37

This work was supported by the National Natural Science Foundation of China NSFC-11101257.

Wei Hong Yang
School of Mathematical Sciences, Fudan University, Shanghai, 200433, P. R. China.
E-mail: whyang@fudan.edu.cn

Lei-Hong Zhang
Department of Applied Mathematics, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, People's Republic of China.
E-mail: longzlh@gmail.com

Ruyi Song
School of Mathematical Sciences, Fudan University, Shanghai, 200433, P. R. China.
E-mail: 09210180023@fudan.edu.cn

1 Introduction

In recent years, there is an increasing interest on the topic of the *optimization methods on manifolds*. Let $\mathscr{M}$ be a *Riemannian manifold* and f be a real-valued function defined on $\mathscr{M}$. Consider the following optimization problem:

$$\min_{x\in\mathscr{M}} f(x). \tag{1.1}$$

If $\mathscr{M}$ is an embedded submanifold of $\mathbb{R}^m$ described by means of smooth equality constraints, (1.1) is usually viewed as a constrained optimization where $\mathscr{M}$ is the feasible region. However, if making full use of the underlying differential geometry of $\mathscr{M}$, we prefer to regard (1.1) as an unconstrained optimization, namely minimizing the function $f:\mathscr{M}\to\mathbb{R}$. To date, many classical methods for the traditional unconstrained minimization, such as the gradient-related algorithms (see Edelman et al. (1998), Gabay (1982), Yang (2007)), the Newton-type methods (see Adler (2002), Mahony (1996), Zhang (2010)) and the trust-region method (see Absil et al. (2007)), have been successfully generalized to the problems in the form of (1.1). The recent monograph by Absil, Mahony and Sepulchre (2008) discusses, in a systematic way, the framework and many numerical first-order and second-order manifold-based algorithms for (1.1), with an emphasis on the application to numerical linear algebra.

However, a lot of optimization problems arising from real-world applications *cannot* simply be formulated as the form of (1.1). In many cases, besides the manifold constraint $x\in\mathscr{M}$, the variable x should also be subject to other equality and/or inequality constraints. For example, in the quadratic assignment problem (see Wen and Yin (2011)), the constraint for the variable X is in the *Stiefel manifold*

$$St(j,k) := \{X\in\mathbb{R}^{k\times j} | X^\top X = I_j\}, \quad (j<k),$$

where I_j stands for the j-by-j identity matrix and also $X\geq 0$ (i.e., each element of X is nonnegative), while in the sparse principal component analysis (see Lu and Zhang (2012)), the variable $X\in St(j,k)$ should also satisfy a set of inequality constraints. Moreover, even if there are only equality constraints, the feasible set in many cases does not form a smooth and connected manifold. This suggests that a more general formula of the nonlinear programming problem should be cast as

$$\begin{aligned} (\mathscr{NLPM}) \qquad \min \quad & f(x) \\ s.t. \quad & c_i(x)=0, \quad i\in\mathscr{E}=\{1,\cdots,l\} \\ & c_i(x)\geq 0, \quad i\in\mathscr{I}=\{l+1,\cdots,r\} \\ & x\in\mathscr{M}. \end{aligned} \tag{1.2}$$

In (1.2), the constraint $x\in\mathscr{M}$ means that both f and c_i for all $i\in\mathscr{E}\cup\mathscr{I}$ are defined on the manifold $\mathscr{M}$.

On the other hand, many real-world problems in the form of (1.2) can also be formulated back into the traditional nonlinear programming problem. This is the case when the manifold $\mathscr{M}$ is an embedded submanifold of $\mathbb{R}^m$. A typical and very important example is again the Stiefel manifold, where the problem (1.1) with $\mathscr{M}=St(j,k)$ can be thought as the equality constrained minimization, and thus the manifold-based algorithms are thereby alternative but efficient approaches for the related equality constrained minimization. It has been observed that when the underlying manifold $\mathscr{M}$ is of simple or nice

differential geometry structure, the manifold-based algorithms appear to be more convenient and can perform better than many state-of-the-art traditional optimization methods (see e.g., Absil (2007, 2008, 2002, 2009)). This is one of our motivations of this paper.

Just as we can generalize the classical numerical algorithms for the unconstrained minimization to (1.2), we can also generalize traditional approaches for the constrained optimization problems and develop counterpart algorithms for (1.2). The idea seems attractive and practical since we have already had several efficient algorithms for (1.1) at hand. In fact, in Wen and Yin (2010), the traditional augmented Lagrangian method is generalized and tried to solve the quadratic assignment problem. This is the other motivation of this paper. Based on our observation, the foremost problem we are facing is the optimality conditions of (1.2). Although discussions on the optimality conditions for (1.1) have been made, e.g., in Absil et al. (2008), Ledyaev and Zhu (2007), Udriste (1988), to the best of the authors' knowledge, systematic treatment on the optimality conditions that combine nonsmooth optimization on a manifold with equality and inequality constraints has not been fully developed. It should be pointed out that the paper Ledyaev and Zhu (2007) discusses this issue based on an appropriate Fréchet subderivative notion, which allows one to develop a quite satisfactory subdifferential calculus. However, we notice that the computation of the subdifferential is difficult according to the definition in Ledyaev and Zhu (2007), and therefore, we will define the subdifferential and treat the optimality conditions in another way, which is of advantage for practical computation.

In this paper, we will discuss optimality conditions for (1.2), under the assumption that $\mathscr{M}$ is a general Riemannian manifold equipped with its *Riemannian connection*, and the constrained functions c_i for all $i \in \mathscr{E} \cup \mathscr{I}$ are differentiable. We will establish the first-order optimality condition of (1.2) when f is a nonsmooth Lipschitz function or continuously differentiable, and present the first-order and the second-order optimality conditions when f is twice continuously differentiable.

We organize the paper in the following way. In Section 2, we will provide some preliminary concepts and notation. In Section 3, we present the definitions of the Clarke generalized gradient of a Lipschitz function on the Riemannian manifolds, the tangent and normal cones to closed subsets of Riemannian manifolds. Then the first-order necessary optimality condition is obtained. In Section 4, we discuss the first-order and the second-order optimality conditions for (1.2) when the objective function is differentiable. In Section 5, we will restrict ourselves to the special case when $\mathscr{M}$ is an embedded submanifold of $\mathbb{R}^m$: For the case that $f(x)$ is nonsmooth Lipschitz continuous, we will study the properties of the Clarke generalized gradient; while more specially, when $\mathscr{M}$ is formed by a set of equality constraints, we show that the optimality conditions for (1.2) can be derived directly from the traditional results on $\mathbb{R}^m$. Final conclusion is draw in Section 6.

2 Preliminaries and Notation

To begin with, we first introduce some basic differential geometry concepts and notations. The reader can find most of these preparatory materials from the books Absil et al. (2008), Lee (2003), Klingenberg (1982).

Let $\mathscr{M}$ be an n-dimensional smooth manifold and $p \in \mathscr{M}$. We will use $\mathfrak{F}_p(\mathscr{M})$ to denote the set of all smooth real-valued functions defined on a neighborhood of p. The *tangent space* to $\mathscr{M}$ at p is represented by $T_p\mathscr{M}$, whose element, known as the *tangent vector*, can be regarded as a mapping ξ_p from $\mathfrak{F}_p(\mathscr{M})$ to $\mathbb{R}$ such that there exists a curve γ on $\mathscr{M}$ with $\gamma(0) = p$, satisfying

$$\xi_p f := \left.\frac{d(f(\gamma(t)))}{dt}\right|_{t=0}$$

for all $f \in \mathfrak{F}_p(\mathscr{M})$. In particular, the tangent vectors $\{\partial/\partial x^i|_p, 1 \le i \le n\}$ defined by

$$\left.\frac{\partial}{\partial x^i}\right|_p f = \left.\frac{\partial(f \circ \varphi^{-1})}{\partial x^i}\right|_{\varphi(p)}, \quad \forall f \in \mathfrak{F}_p(\mathscr{M}) \tag{2.1}$$

forms a basis of $T_p\mathscr{M}$. The tangent bundle $T\mathscr{M} := \cup_p T_p\mathscr{M}$ consists of all tangent vectors to $\mathscr{M}$.

Let $F : \mathscr{M} \to \mathscr{N}$ be a smooth mapping between two smooth manifolds $\mathscr{M}$ and $\mathscr{N}$. The *differential* (also known as *push-forward*) of F at p (see Absil et al. (2008)), $\mathbf{D}F(p) : T_p\mathscr{M} \to T_{F(p)}\mathscr{N}$ is defined by

$$(\mathbf{D}F(p)[\xi_p])f := \xi_p(f \circ F),$$

for all $\xi_p \in T_p\mathscr{M}$ and $f \in \mathfrak{F}_{F(p)}(\mathscr{N})$. By (2.1), we have $\mathbf{D}\varphi(p)\frac{\partial}{\partial x^i}|_p = \frac{\partial}{\partial x^i}|_{\varphi(p)}$ and therefore, if $\xi = \sum_{i=1}^n \xi_i \frac{\partial}{\partial x^i}|_p \in T_p\mathscr{M}$, we have

$$\mathbf{D}\varphi(p)[\xi] = (\xi_1, \cdots, \xi_n)^\top \in \mathbb{R}^n.$$

For notational convenience thus, we will use $\hat{\bullet}$ to represent the corresponding counterpart of the object $\bullet$ related with $\mathscr{M}$ in $\mathbb{R}^n$:

$$\hat{p} := \varphi(p), \;\; \hat{\xi} := \mathbf{D}\varphi(p)[\xi], \;\; \text{and} \;\; \hat{f} := f \circ \varphi^{-1}. \tag{2.2}$$

With this notation, for $\xi_p \in T_p\mathscr{M}$ and $f \in \mathfrak{F}_p(\mathscr{M})$, it is easy to see that

$$\xi_p f = \langle \hat{\xi}_p, \nabla \hat{f}(\hat{p}) \rangle.$$

A differentiable manifold whose tangent spaces are endowed with a smoothly varying inner product with respect to $p \in \mathscr{M}$ is called a Riemannian manifold. The smoothly varying inner product, denoted by $\langle \cdot, \cdot \rangle_p$, is called the *Riemannian metric*, and when no confusion arises, we will also omit the subscript and use simply $\langle \cdot, \cdot \rangle$ instead of $\langle \cdot, \cdot \rangle_p$. Let $g_{ij}(p) := \langle \partial/\partial x^i|_p, \partial/\partial x^j|_p \rangle_p$. Then $g_{ij}(\cdot)$ is a smooth function on $\mathscr{M}$. Thus, for vector fields $\xi = \sum_i \xi_i \partial/\partial x^i$ and $\zeta = \sum_i \zeta_i \partial/\partial x^i$, we have

$$g(\xi, \zeta) := \langle \xi, \zeta \rangle = \sum_{ij} g_{ij} \xi_i \zeta_j.$$

Thus, if we introduce the notation $G : \hat{p} \mapsto G_{\hat{p}}$ to denote the matrix-valued function such that the (i, j) element of $G_{\hat{p}}$ is $g_{ij}(p)$, we have then, in matrix notation,

$$g(\xi, \zeta) = \langle \xi, \zeta \rangle = \hat{\xi}^\top G_{\hat{p}} \hat{\zeta}. \tag{2.3}$$

In our discussion on the optimality condition for (1.2), we will restrict ourselves to the case that $\mathscr{M}$ is a Riemannian manifold with a Riemannian metric g.

Given $f \in \mathfrak{F}_p(\mathscr{M})$, the *gradient* of f at p (see Absil et al. (2008)), denoted by $\mathrm{grad} f(p)$, is defined as the unique tangent vector in $T_p\mathscr{M}$ that satisfies the condition:

$$\langle \mathrm{grad} f(p), \xi \rangle := \xi_p f = \langle \nabla \hat{f}(\hat{p}), \hat{\xi} \rangle, \quad \forall \xi \in T_p\mathscr{M}, \tag{2.4}$$

and so in matrix notation, the coordinate expression of $\mathrm{grad} f(p)$ is given by (see Absil et al. (2008))

$$\mathbf{D}\varphi(p)[\mathrm{grad} f(p)] = G_{\hat{p}}^{-1} \nabla \hat{f}(\hat{p}). \tag{2.5}$$

The *length* of a curve $\gamma : [a,b] \to \mathscr{M}$ on $\mathscr{M}$ is defined by

$$L(\gamma) = \int_a^b \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle} dt,$$

and the *Riemannian distance* (see Absil et al. (2008)) on $\mathscr{M}$ is given by

$$d : \mathscr{M} \times \mathscr{M} \to \mathbb{R} : \ \ d(x,y) = \inf_{\Gamma} L(\gamma),$$

where Γ represents the set of all curves in $\mathscr{M}$ joining points x and y. Thus the set $\{y \in \mathscr{M} \mid d(x,y) < \delta\}$ serves as a neighborhood of x with radius $\delta > 0$.

A *geodesic* is a curve on $\mathscr{M}$ that locally minimizes the arc length. For every $\xi \in T_p\mathscr{M}$, there exists an interval $\mathscr{Z}$ containing 0 and a unique geodesic $\gamma(t; p, \xi) : \mathscr{Z} \to \mathscr{M}$ such that $\gamma(0) = p$ and $\dot{\gamma}(0) = \xi$. The mapping (see Absil et al. (2008))

$$Exp_p : T_p\mathscr{M} \to \mathscr{M} : \xi \mapsto Exp_p \xi = \gamma(1; p, \xi)$$

is called the *exponential map* on $p \in \mathscr{M}$. Let $E : \mathbb{R}^n \to T_p\mathscr{M}$ be a linear bijection such that $\{E(e_i)\}_{i=1}^n$ is an orthogonal basis for $T_p\mathscr{M}$, where e_i is the i-th unit vector. Let U be a neighborhood of a point p and V a neighborhood of the origin of $T_p\mathscr{M}$ such that Exp_p is a diffeomorphism between V and U. If we define

$$\varphi = E^{-1} Exp_p^{-1}. \tag{2.6}$$

then (U, φ) is known as a *Riemannian normal coordinate system*. Under the normal coordinate system, it is true that (see Klingenberg (1982))

$$G_{\hat{p}} = I_n. \tag{2.7}$$

Lastly, the *Riemannian Hessian* (see Absil et al. (2008)) of $f \in \mathfrak{F}_p(\mathscr{M})$ at a point p in $\mathscr{M}$ is defined as the (symmetric) linear mapping $\mathrm{Hess} f(p)$ of $T_p\mathscr{M}$ into itself that satisfies

$$\mathrm{Hess} f(p)[\xi] = \nabla_{\xi} \mathrm{grad} f, \qquad \forall \xi \in T_p\mathscr{M},$$

where ∇ stands for the *Riemannian connection* (see Absil et al. (2008)) on $\mathscr{M}$. The following result (see Prop. 5.5.4 of Absil et al. (2008)) is useful for our analysis

$$\mathrm{Hess} f(p) = \mathrm{Hess}(f \circ Exp_p)(0_p). \tag{2.8}$$

3 The Clarke generalized gradient

In this section, we will study the property of a Lipschitz function definded on a Riemannian manifold $\mathscr{M}$. The Lipschitz behavior of functions on $\mathscr{M}$ has been studied in Ferreira (2006, 2008). In this paper, we only focus on the Clarke generalized gradient of a Lipschitz function and its application to the optimality condition.

3.1 Lipschitz continuity

Recall that a function f on $\mathscr{M}$ is said to be Lipschitz of rank L on a set U if

$$|f(y)-f(z)| \leq L \cdot d(y,z), \quad \forall y,z \in U,$$

where $L>0$. If there exists a neighborhood U of $p \in \mathscr{M}$ such that f is Lipschitz of rank L on U, we say that f is Lipschitz of rank L at p; if furthermore, for every $p \in \mathscr{M}$, f is Lipschitz of rank L at p for some $L>0$, then f is said to be locally Lipschitz on $\mathscr{M}$.

Now suppose $f:\mathscr{M}\to\mathbb{R}$ is a locally Lipschitz function on $\mathscr{M}$. The *generalized directional derivative* of f at $p\in\mathscr{M}$ in the direction $v\in T_p\mathscr{M}$, denoted by $f^\circ(p;v)$, is defined as (see Hosseini and Pouryayevali (2011))

$$f^\circ(p;v) := \limsup_{y\to p, t\downarrow 0} \frac{f\circ\varphi^{-1}(\varphi(y)+t\mathbf{D}\varphi(p)(v)) - f\circ\varphi^{-1}(\varphi(y))}{t}, \tag{3.1}$$

where (U,φ) is a chart containing p. Indeed, we have

$$f^\circ(p;v) = (f\circ\varphi^{-1})^\circ(\varphi(p);\mathbf{D}\varphi(p)(v)), \tag{3.2}$$

where $(f\circ\varphi^{-1})^\circ(\varphi(p);\mathbf{D}\varphi(p)(v))$ is the Clark generalized directional derivative (see Clarke (1983)) of $f\circ\varphi^{-1}$ at $\varphi(p)$. It should be pointed out that this definition does not depend on the choice of charts (see Motreanu and Pavel (1982)).

Lemma 3.1 *Let $\mathscr{M}$ be a Riemannian manifold and $p\in\mathscr{M}$. Let (U,φ) be a chart at p. Then f is Lipschitz at p if and only if $f\circ\varphi^{-1}$ is Lipschitz at $\varphi(p)$.*

Proof Let (U,φ) and (V,ψ) be two charts containing p. Since the Jacobian matrix of the mapping $\varphi\circ\psi^{-1}$ is nonsingular around $\psi(p)$, we only need to prove the assertion for one *particular* chart (U,φ), namely, the normal coordinate defined by (2.6). We use the notation $B_\delta := \{y\in\mathscr{M} : d(p,y)\leq\delta\}$ and $B(0_p;\delta) := \{\xi\in T_p\mathscr{M} : \|\xi\|\leq\delta\}$. By Theorem 2.3 in Azagra et al (2005), for every $C>1$, the mappings $Exp_p : B(0_p;\delta)\to B_\delta$ and $Exp_p^{-1} : B_\delta \to B(0_p;\delta)$ are C-Lipschitz for small enough δ, and so the assertion follows.

The proofs of the following results are similar to Proposition 2.1.1 in Clarke (1983) except for (3.3). Thus, we only give the proof for (3.3).

Theorem 3.1 *Let $\mathscr{M}$ be a Riemannian manifold. Suppose that the function $f:\mathscr{M}\to\mathbb{R}$ is Lipschitz of rank L on an open set V. Then,*

(i) *for each $p \in V$ and $v \in T_p\mathscr{M}$, the function $v \mapsto f^\circ(p;v)$ is finite, positive homogeneous, and subadditive on $T_p\mathscr{M}$, and satisfies*

$$|f^\circ(p;v)| \le L\|v\|; \tag{3.3}$$

(ii) *$f^\circ(p;v)$ is upper semicontinuous on $TM|_U$ and, as a function of v alone, is Lipschitz of rank L on $T_p\mathscr{M}$, for each $p \in U$;*

(iii) *$f^\circ(p;-v) = (-f)^\circ(p;v)$ for each $p \in U$ and $v \in T_p\mathscr{M}$.*

Proof Since $f^\circ(p;v)$ does not depend on chart, we consider the normal coordinate (U,φ) defined by (2.6). Assume that $U \subset V$. Write $v \in T_p\mathscr{M}$ as $v = \sum_{i=1}^n v_i \frac{\partial}{\partial x^i}|_p$. Let $\hat{v} = (v_1,\cdots,v_n)^\top \in \mathbb{R}^n$. By (2.7), we have $G_{\varphi(p)} = I_n$, and so $\|v\| = \sqrt{\langle v,v\rangle_p} = \sqrt{\hat{v}^\top G_{\varphi(p)}\hat{v}} = \|\hat{v}\|$.

We use the notation $U_\delta := \{y \in \mathscr{M} : d(p,y) < \delta\}$. Fix $\varepsilon > 0$. Since $G_{\varphi(y)}$ is a continuous functions of y, there exists δ_0 such that $U_{\delta_0} \subset U$ and if $y \in U_{\delta_0}$, then the largest eigenvalue $\lambda_{max}(G_{\varphi(y)})$ of $G_{\varphi(y)}$ satisfies

$$\lambda_{max}(G_{\varphi(y)}) < 1 + 2\varepsilon.$$

Denote $\eta_t := \varphi(y) + t\hat{v} \in \mathbb{R}^n$ and $y(t) := \varphi^{-1}(\eta_t)$. Then $\dot{y}(t) = \sum v_i \frac{\partial}{\partial x^i}|_{y(t)}$, and so

$$d(y,y(t)) \le \int_0^t \sqrt{\langle \dot{y}(s),\dot{y}(s)\rangle}ds = \int_0^t \sqrt{\hat{v}^\top G_{\eta_s}\hat{v}}ds \le (1+\varepsilon)t\|\hat{v}\| = (1+\varepsilon)t\|v\|. \tag{3.4}$$

From definition (3.1), we know that there exist δ_1 and t_0 such that

$$f^\circ(p;v) \le \sup_{y\in U_{\delta_1}, 0<t<t_0} \frac{f(y(t)) - f(y)}{t} + \varepsilon,$$

which implies that there exist $y \in U_{\delta_1}$ and $t \in (0,t_0)$ such that

$$f^\circ(p;v) \le \frac{f(y(t)) - f(y)}{t} + 2\varepsilon. \tag{3.5}$$

Let δ_1 and t_0 be both small enough so that (3.4) holds, then we have from (3.5)

$$f^\circ(p;v) \le L\frac{d(y,y(t))}{t} + 2\varepsilon \le L(1+\varepsilon)\|v\| + 2\varepsilon.$$

Since ε is arbitrary, (3.3) holds.

The generalized gradient or the *Clarke subdifferential* of a locally Lipschitz function f at $p \in \mathscr{M}$, denoted by $\partial f(p)$, is defined as

$$\partial f(p) = \{\xi \in T_p\mathscr{M} : \langle \xi, v\rangle \le f^\circ(p;v) \text{ for all } v \in T_p\mathscr{M}\}. \tag{3.6}$$

By (2.3), (3.2) and (3.6), we have the following result.

Proposition 3.1 *Let $(\mathscr{M},g)$ be a Riemannian manifold and $p \in \mathscr{M}$. Suppose that $f : \mathscr{M} \to \mathbb{R}$ is Lipschitz near p and (U,φ) is a chart at p. Then*

$$\partial f(p) = [\boldsymbol{D}\varphi(p)]^{-1}[G^{-1}_{\varphi(p)}\partial(f\circ\varphi^{-1})(\varphi(p))].$$

Remark 3.1 In Hosseini and Pouryayevali (2011), the generalized gradient $\partial f(p)$ is defined on $T_p^*\mathscr{M}$, the cotangent space at p, and satisfies the property (see Proposition 2.5 in Hosseini and Pouryayevali (2011)):

$$\partial f(p) = \mathbf{D}\varphi(p)^*[\partial(f\circ\varphi^{-1})(\varphi(p))],$$

where $*$ denotes the adjoint. Since the gradient of a differentiable function is defined on the tangent space, from the computational point of view, we think it is more reasonable to define the generalized gradient of a nonsmooth function (a generalization of gradient) on the tangent space as in (3.6).

Similar to Theorem 2.9 in Hosseini and Pouryayevali (2011), it is easy to prove the following results and we omit the proof.

Theorem 3.2 *Let $\mathscr{M}$ be a Riemannian manifold, $p\in\mathscr{M}$ and f is Lipschitz of rank L on some neighborhood U of p. Then,*

(i) *$\partial f(p)$ is a nonempty, convex, compact subset of $T_p\mathscr{M}$, and $\|\xi\|\le L$ for every $\xi\in\partial f(p)$;*
(ii) *for every $v\in T_p\mathscr{M}$, we have*

$$f^\circ(p;v) = \max\{\langle\xi,v\rangle : \xi\in\partial f(p)\},$$

and so $\xi\in\partial f(p)$ if and only if $f^\circ(p;v)\ge\langle\xi,v\rangle$;
(iii) *let $\{p_i\}$ and $\{\xi_i\}$ be sequences in $\mathscr{M}$ and $T\mathscr{M}$ such that $\xi_i\in\partial f(p_i)$. Suppose that p_i converges to p and $\eta\in\mathbb{R}^n$ is a cluster point of the $G_{\varphi(x_i)}[\boldsymbol{D}\varphi(p_i)\xi_i]$, then we have $[\boldsymbol{D}\varphi(p)]^{-1}[G^{-1}_{\varphi(p)}\eta]\in\partial f(p)$.*

3.2 Tangent and normal cone

Let C be a subset in $\mathbb{R}^n$ and $x\in C$. We use $T_C(x)$ (resp. $N_C(x)$) to denote the *(Clarke) tangent (resp. normal) cone* (see Clarke (1983)) to C at x. By Theorem 2.4.5 in Clarke, the (Clarke) tangent cone is in accordance with the one defined by Definition 12.2 in Nocedal and Wright (2006).

By making use of the coordinate chart, we can define the tangent cone of a nonempty closed subset of Riemannian manifold $\mathscr{M}$. To this end, assume S is a nonempty closed subset of Riemannian manifold $\mathscr{M}$ and $p\in S$, and (U,φ) is a chart of $\mathscr{M}$ at p. Then we define the *(Clarke) tangent cone* $\mathscr{T}_S(p)$ to S at p as follows:

$$\mathscr{T}_S(p) := [\mathbf{D}\varphi(p)]^{-1}[T_{\varphi(S\cap U)}(\varphi(p))]. \tag{3.7}$$

It is true that this definition of $\mathscr{T}_S(p)$ does not depend on the choice of the chart (U,φ) at p (see Motreanu and Pavel (1982)). Furthermore, we can also define the normal cone, denoted by $\mathscr{N}_S(p)$, to S at p as

$$\mathscr{N}_S(p) := \{u\in T_p\mathscr{M} : \langle u,v\rangle\le 0,\ \ \forall v\in\mathscr{T}_S(p)\}. \tag{3.8}$$

Using (2.3) and (3.8), it is easy to prove the following result.

Proposition 3.2 *We have $\mathscr{N}_S(p) = [\boldsymbol{D}\varphi(p)]^{-1}[G^{-1}_{\varphi(p)}N_{\varphi(S\cap U)}(\varphi(p))]$.*

For a function f defined on a set $S \subseteq \mathscr{M}$, we say that f attains a local minimum over S at p if there exists a neighborhood $V \subseteq \mathscr{M}$ of p such that $f(y) \geq f(p)$, $\forall y \in V \cap S$.

Proposition 3.3 *Suppose that f is Lipschitz at p and attains a local minimum over a set S at p. Then* $0 \in \partial f(p) + \mathscr{N}_S(p)$.

Proof Let (U, φ) be a chart around p. By Lemma 3.1, $\hat{f} = f \circ \varphi^{-1}$ is also Lipshitz and $\hat{f}$ attains a local minimum over the set $\varphi(S \cap U)$ at $\hat{p}$. From the proof on page 52 in Clarke (1983), we have $0 \in \partial \hat{f}(\hat{p}) + N_{\varphi(S\cap U)}(\hat{p})$, which together with Propositions 3.1 and 3.2 implies the assertion.

Remark 3.2 In Hosseini and Pouryayevali (2011), the normal cone is again defined on the cotangent space. For our discussion, according to Proposition 3.3, we prefer the definition (3.8) because $\partial f(p) \subseteq T_p\mathscr{M}$. In particular, if f is continuously differentiable and attains a minimum over a set $S \subset \mathscr{M}$ at x, then similar to the proof in Proposition 3.3, it is true that

$$-\text{grad} f(x) \in \mathscr{N}_S(x),$$

which is a generalization of the traditional result on $\mathbb{R}^n$. Since the gradient is a tangent vector of $\mathscr{M}$, it is more reasonable to define the normal cone on the tangent space.

4 Necessary Optimality Conditions for Constrained Problems

4.1 First-order optimality conditions

Now we consider the problem of the form (1.2) in which f is a locally Lipschitz function and c_i for all $i \in \mathscr{E} \cup \mathscr{I}$ are differentiable functions. We denote the feasible region of (1.2) by Ω.

Using coordinate charts, we can transform (1.2) into a traditional nonlinear programming problem locally. Indeed, for $x \in \Omega$, suppose (U, φ) is a chart around x, then we have the following nonlinear programming problem in $\mathbb{R}^n$:

$$\begin{aligned} \min \quad & \hat{f}(\hat{x}) \\ s.t. \quad & \hat{c}_i(\hat{x}) = 0, \quad i \in \mathscr{E} = \{1, \cdots, l\} \\ & \hat{c}_i(\hat{x}) \geq 0, \quad i \in \mathscr{I} = \{l+1, \cdots, r\} \\ & \hat{x} \in \varphi(U) \subseteq \mathbb{R}^n, \end{aligned} \tag{4.1}$$

where $\hat{f}$ and $\hat{c}_i$ are defined by (2.2). If we define the active set $\mathscr{A}(x)$ at $x \in \mathscr{M}$ of (1.2) by

$$\mathscr{A}(x) := \mathscr{E} \cup \{i \in \mathscr{I} \mid c_i(x) = 0\}, \tag{4.2}$$

then it is clear that $A(\hat{x}) = \mathscr{A}(x)$, where $A(\hat{x})$ is the active set of $\hat{x}$ for the problem (4.1). With the active set $\mathscr{A}(x)$, we say that the *linear independent constraint qualification* (LICQ) on manifold $\mathscr{M}$ holds at x if

$$\{\text{grad } c_i(x), i \in \mathscr{A}(x)\} \text{ is linear independent on } T_x\mathscr{M}. \tag{4.3}$$

It is not difficult to check by (2.5) that (4.3) is equivalent to the statement that $\{\nabla \hat{c}_i(\hat{x}), i \in \mathscr{A}(x)\}$ is linear independent, that is $\{\nabla \hat{c}_i(\hat{x}), i \in \mathscr{A}(x)\}$ satisfies the LICQ (see Nocedal and Wright (2006)) on $\mathbb{R}^n$. To further explore the optimality condition, we next define the concept of linearized feasible direction.

Definition 4.1 Given a feasible point $x \in \mathscr{M}$ and the active constraint set $\mathscr{A}(x)$ given by (4.2), the set of linearized feasible directions $\mathscr{F}(x)$ is defined by

$$\mathscr{F}(x) = \left\{ d \in T_x\mathscr{M} \;\middle|\; \begin{array}{ll} \langle \text{grad } c_i(x), d\rangle = 0, & \text{for all } i \in \mathscr{E}, \\ \langle \text{grad } c_i(x), d\rangle \geq 0, & \text{for all } i \in \mathscr{A}(x) \cap \mathscr{I} \end{array} \right\}.$$

For any $d \in \mathscr{F}(x)$, since $\langle \text{grad } c_i(x), d\rangle = \langle \nabla \hat{c}_i(\hat{x}), \hat{d}\rangle$, we have

$$\hat{d} \in F(\hat{x}), \tag{4.4}$$

where $F(\hat{x})$ is the set of linearized feasible directions of the problem (4.1) (see Definition 12.3 in Nocedal and Wright (2006)) given by

$$F(\hat{x}) = \left\{ \hat{d} \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} \langle \nabla \hat{c}_i(x), \hat{d}\rangle = 0, & \text{for all } i \in \mathscr{E}, \\ \langle \nabla \hat{c}_i(x), \hat{d}\rangle \geq 0, & \text{for all } i \in A(\hat{x}) \cap \mathscr{I} \end{array} \right\}. \tag{4.5}$$

By Lemma 12.2 in Nocedal and Wright (2006), (4.4) and (3.7), it is straightforward to prove the following lemma.

Lemma 4.1 *Let $x \in \Omega$ be a feasible point. The following two statements are true:*

(i) $\mathscr{T}_\Omega x \subseteq \mathscr{F}(x)$, *and*
(ii) *if the LICQ condition* (4.3) *is satisfied at x, then* $\mathscr{T}_\Omega x = \mathscr{F}(x)$.

With Lemma 4.1 (ii) in hand, following the procedure of proving Lemma 12.9 in Nocedal and Wright (2006), we can characterize the normal cone $\mathscr{N}_\Omega(x)$ in the following way:

Corollary 4.1 *If the LICQ condition* (4.3) *is satisfied at x, then*

$$\mathscr{N}_\Omega(x) = \{ \sum_{i \in \mathscr{A}(x)} \lambda_i \text{grad } c_i(x), \quad \lambda_i \leq 0 \text{ for } i \in \mathscr{A}(x) \cap \mathscr{I} \}. \tag{4.6}$$

By Proposition 3.3 and Corollary 4.1 consequently, it is easy to prove the following result.

Theorem 4.1 (First-Order Necessary Conditions) *Suppose that x^* is a local solution of* (1.2) *and that the LICQ* (4.3) *holds at x^*. Define the Lagrangian function for problem* (1.2) *as*

$$\mathscr{L}(x, \lambda) = f(x) - \sum_{i \in \mathscr{E} \cup \mathscr{I}} \lambda_i c_i(x). \tag{4.7}$$

Then there is a Lagrange multiplier vector λ^*, with components λ_i^*, $i \in \mathscr{E} \cup \mathscr{I}$, such that the following conditions are satisfied at (x^*, λ^*)

$$\begin{aligned} 0 \in \partial_x \mathscr{L}(x^*, \lambda^*), & \\ c_i(x^*) = 0, \quad & \text{for all } i \in \mathscr{E}, \\ c_i(x^*) \geq 0, \quad \lambda_i^* \geq 0, \quad \lambda_i^* c_i(x^*) = 0, \quad & \text{for all } i \in \mathscr{I}. \end{aligned} \tag{4.8}$$

Remark 4.1 Consider the nonlinear programming problem (4.1). The Lagrangian function of (4.1) is $L(\hat{x},\lambda) = \hat{f}(\hat{x}) - \sum_{i\in\mathscr{E}\cup\mathscr{I}} \lambda_i \hat{c}_i(\hat{x})$. Let λ^* be the Lagrange multiplier satisfying (4.8). By Proposition 3.1, it is easy to prove that $0 \in \partial_{\hat{x}} L(\hat{x}^*,\lambda^*)$, which means that λ^* is also the Lagrange multiplier of (4.1).

Remark 4.2

1. In the case that f is continuously differentiable, replacing $0 \in \partial_x \mathscr{L}(x^*,\lambda^*)$ in (4.8) by $\mathrm{grad}_x \mathscr{L}(x^*,\lambda^*) = 0$, we get the first-order necessary condition of (1.2).
2. More particularly, if f is a continuously differentiable function and x^* is a local solution of $\min_{x\in\mathscr{M}} f(x)$, then we have $\mathrm{grad} f(x^*) = 0$.

4.2 Second-order optimality conditions

Now we are in a position to describe the second-order optimality conditions for (1.2). In this subsection, assume that f and c_i, $i \in \mathscr{E}\cup\mathscr{I}$, are twice differentiable. Suppose x^* is a local solution of (1.2) and λ^* is some Lagrange multiplier vector that satisfies the KKT conditions (4.8). We further define the *critical cone* $\mathscr{C}(x^*,\lambda^*)$ associated with (x^*,λ^*) as follows:

$$\begin{aligned} &\mathscr{C}(x^*,\lambda^*) \\ &= \{w \in \mathscr{T}(x^*) \mid \langle \mathrm{grad}\, c_i(x^*), w\rangle = 0, \text{ for all } i \in \mathscr{A}(x^*)\cap\mathscr{I} \text{ with } \lambda_i^* > 0\}. \end{aligned} \tag{4.9}$$

Equivalently, one can readily verify that

$$w \in \mathscr{C}(x^*,\lambda^*) \Leftrightarrow \begin{cases} w \in T_{x^*}\mathscr{M}, \\ \langle \mathrm{grad}\, c_i(x^*), w\rangle = 0, & \text{for all } i \in \mathscr{E}, \\ \langle \mathrm{grad}\, c_i(x^*), w\rangle = 0, & \text{for all } i \in \mathscr{A}(x^*)\cap\mathscr{I} \text{ with } \lambda_i^* > 0, \\ \langle \mathrm{grad}\, c_i(x^*), w\rangle \geq 0, & \text{for all } i \in \mathscr{A}(x^*)\cap\mathscr{I} \text{ with } \lambda_i^* = 0. \end{cases}$$

Given $F(\hat{x}^*)$ defined by (4.5) and λ^* satisfying (4.8), let the critical cone (cf. page 330 in Nocedal and Wright (2006)) $C(\hat{x^*},\lambda^*)$ be defined by

$$\hat{w} \in C(\hat{x}^*,\lambda^*) \Leftrightarrow \begin{cases} \langle \nabla \hat{c}_i(x^*), \hat{w}\rangle = 0, & \text{for all } i \in \mathscr{E}, \\ \langle \nabla \hat{c}_i(x^*), \hat{w}\rangle = 0, & \text{for all } i \in \mathscr{A}(x^*)\cap\mathscr{I} \text{ with } \lambda_i^* > 0, \\ \langle \nabla \hat{c}_i(x^*), \hat{w}\rangle \geq 0, & \text{for all } i \in \mathscr{A}(x^*)\cap\mathscr{I} \text{ with } \lambda_i^* = 0. \end{cases}$$

By (2.4), we have

$$w \in \mathscr{C}(x^*,\lambda^*) \Longleftrightarrow \hat{w} \in C(\hat{x}^*,\lambda^*). \tag{4.10}$$

With the aid of these results, we can state the second-order necessary optimality conditions as follows:

Theorem 4.2 (Second-Order Necessary Conditions) *Suppose that x^* is a local solution of* (1.2) *and that the LICQ condition* (4.3) *is satisfied. Let λ^* be the Lagrange multiplier vector for which* (4.8) *are satisfied. Then*

$$\mathrm{Hess}_x \mathscr{L}(x^*,\lambda^*)[w,w] \geq 0, \quad \forall w \in \mathscr{C}(x^*,\lambda^*).$$

Proof Let (U,φ) be the normal coordinate defined by (2.6). Consider the nonlinear programming problem (4.1). It is obvious that $\hat{x}^*$ is a local solution of (4.1), and λ^* is also the Lagrange multiplier vector for which the KKT conditions hold. Then the assertion follows from (2.8), (4.10) and Theorem 12.5 in Nocedal and Wright (2006).

The following corollary is a direct application of Theorem 4.2 for the case $\mathscr{E}=\mathscr{I}=\emptyset$.

Corollary 4.2 *Let x^* be a local solution of* $\min_{x\in\mathscr{M}} f(x)$. *Then* $\mathrm{Hess} f(x^*)$ *is positive semidefinite on $T_{x^*}\mathscr{M}$.*

Similar to the proof of Theorem 4.2, we can also establish the following second-order sufficient conditions for problem (1.2).

Theorem 4.3 (Second-Order Sufficient Conditions) *Suppose that for some feasible point $x^*\in\mathscr{M}$ there is a Lagrange multiplier vector λ^* such that the KKT conditions* (4.8) *are satisfied. Suppose also that*

$$\mathrm{Hess}_x\mathscr{L}(x^*,\lambda^*)[w,w]>0,\quad \forall w\in\mathscr{C}(x^*,\lambda^*),\ w\neq 0.$$

Then x^* is a strict local solution for (1.2).

Analogously, in the special case $\mathscr{E}=\mathscr{I}=\emptyset$, Theorem 4.3 reduces the following second-order sufficient optimality condition.

Corollary 4.3 *Let $x^*\in\mathscr{M}$ be such that* $\mathrm{grad} f(x^*)=0$ *and* $\mathrm{Hess} f(x^*)$ *is positive definite on $T_{x^*}\mathscr{M}$. Then x^* is a strict local solution of* $\min_{x\in\mathscr{M}} f(x)$.

5 The case when $\mathscr{M}$ is an embedded submanifold of $\mathbb{R}^m$

In Section 4, we have established the optimality conditions for the case when $\mathscr{M}$ is a general n-dimensional Riemannian manifold. In real-world applications, however, it turns out that the manifolds we encounter can always be viewed as embedded submanifolds of $\mathbb{R}^m$ with $m>n$. In this section, we will consider the case that $\mathscr{M}$ is an embedded submanifold (for definition, see Absil et al. (2008)) of $\mathbb{R}^m$. In this case $T_x\mathscr{M}$ is a subspace of $\mathbb{R}^m$ for any $x\in\mathscr{M}$.

We use P_x to denote the orthogonal projection onto $T_x\mathscr{M}$. Then $T_x\mathscr{M}=\{P_x y|y\in\mathbb{R}^m\}$. This expression of the tangent space is very useful, because for any differentiable real-valued function h defined on $\mathbb{R}^m$, the gradient of the restriction $h|_{\mathscr{M}}:\mathscr{M}\to\mathbb{R}$, of h on $\mathscr{M}$ can be simply formulated as (see (3.3.7) in Absil et al. (2008))

$$\mathrm{grad} h|_{\mathscr{M}}(x)=P_x\nabla h(x), \tag{5.1}$$

where $\nabla h(x)$ stands for the gradient of h, viewed as a function defined on $\mathbb{R}^m$, at x.

Let f be a nonsmooth Lipschitz function. We use the notation

$$\bar{f}:=f|_{\mathscr{M}} \tag{5.2}$$

to denote the restriction of f on $\mathscr{M}$. It is easy to prove that $\bar{f}$, as a function on manifold $\mathscr{M}$, is also a Lipschitzian function. Motivated by (5.1), it is natural to ask whether $\partial\bar{f}(x)=P_x\partial f(x)$. As we will see in Theorem 5.1 that $\partial\bar{f}(x)\subset P_x\partial f(x)$ is always true, but the converse is not, as Example 5.1 demonstrates.

Example 5.1 Let $\mathscr{M} = \{(x_1,0) : x_1 \in \mathbb{R}\}$ be the embedded submanifold in $\mathbb{R}^2$. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x_1,x_2) = \begin{cases} -x_2, \text{ if } x_1 \geq -x_2, x_2 \leq 0; \\ x_1, \text{ if } x_1 \leq -x_2, x_1 \geq 0; \\ 0, \text{ otherwise.} \end{cases}$$

Then f equals to 0 on the whole $\mathscr{M}$ and so $\bar{f} \equiv 0$. Let $x = (0,0)$ and $v = (1,0)$. Then $f^\circ(x;v) = 1$ and $\bar{f}^\circ(x;v) = 0$. It is easy to see that $\partial f(x) = \{(x_1,0) : 0 \leq x_1 \leq 1\}$. Thus we have $[0,1] = P_x\partial f(x) \neq \partial \bar{f}(x) = \{0\}$.

We next will establish a sufficient condition for the relation $\partial \bar{f}(x) = P_x\partial f(x)$.

Definition 5.1 A function f is said to be regular at $x \in \mathscr{M}$ along $T_x\mathscr{M}$ if

(i) for all $v \in T_x\mathscr{M}$, $f'(x;v) := \lim_{t\downarrow 0} \frac{f(x+tv)-f(x)}{t}$ exists, and
(ii) for all $v \in T_x\mathscr{M}$, $f'(x;v) = f^\circ(x;v)$.

Theorem 5.1 *Let $\mathscr{M}$ be an embedded submanifold of $\mathbb{R}^m$. Let f be a Lipschitz function at $x \in \mathscr{M}$ and $\bar{f}$ be defined by* (5.2)*. Then we have*

(i) *$f^\circ(x;d) \geq \bar{f}^\circ(x;d)$ for any $d \in T_x\mathscr{M}$,*
(ii) *$\partial \bar{f}(x) \subset P_x\partial f(x)$, and*
(iii) *if f is regular at x along $T_x\mathscr{M}$, then $\partial \bar{f}(x) = P_x\partial f(x)$.*

Proof (i). Pick any $d \in T_x\mathscr{M}$. Let (U,φ) be a normal coordinate system around x. By the definition of the normal coordinate system, we have $\mathbf{D}\varphi(x) = Id_{T_x\mathscr{M}}$, where $Id_{T_x\mathscr{M}}$ denotes the identity mapping on $T_x\mathscr{M}$. Let $U_\delta := \{y \in \mathscr{M} : d(x,y) < \delta\}$. Fix $\varepsilon > 0$. Then there exists $\delta_1 > 0$ such that for all $y \in U_{\delta_1}$,

$$\|\mathbf{D}\varphi^{-1}(\varphi(y))(\mathbf{D}\varphi(x)(u)) - \mathbf{D}\varphi(x)(u)\| \leq \varepsilon\|u\|, \quad \forall u \in T_x\mathscr{M}. \tag{5.3}$$

Let $y(t) = \varphi^{-1}(\varphi(y) + t\mathbf{D}\varphi(x)(d))$. Since φ is a diffeomorphism, we have

$$y(t) = y + t\mathbf{D}\varphi^{-1}(\varphi(y))(\mathbf{D}\varphi(x)(d)) + t^2\psi(y),$$

where ψ is a smooth function of y. Therefore, together with (5.3) and $\mathbf{D}\varphi(x) = Id_{T_x\mathscr{M}}$, it implies that for sufficiently small t

$$\|y(t) - (y + td)\| \leq \varepsilon t\|d\| + Mt^2, \quad \forall y \in U_{\delta_1}, \tag{5.4}$$

where M could be chosen as a constant independent on y. By definition of the Clarke generalized directional derivative, it follows that there are $\delta_2 > 0$ and $t_0 > 0$ such that for any $\delta \in (0,\delta_2)$ and $t \in (0,t_0)$, we have

$$\sup_{\|z-x\|<\delta, 0<s<t} \frac{f(z+sd)-f(z)}{s} - f^\circ(x;d) \leq \varepsilon; \tag{5.5}$$

on the other hand, by definition of (3.1), for any sufficiently small $\delta \in (0,\delta_2)$ and $t \in (0,t_0)$, there exist some $\bar{y} \in U_\delta$ and $0 < \bar{t} < t$ satisfying

$$\bar{f}^\circ(x;d) - \frac{f(y(\bar{t})) - f(\bar{y})}{\bar{t}} \leq \varepsilon. \tag{5.6}$$

Let δ and t be small enough so that the Lipschitz constant of f on $\{y : \|y-x\| \leq \delta\}$ is L and $LMt \leq \varepsilon$. Then by (5.4), we have

$$\Big|\frac{f(y(\bar{t})) - f(\bar{y})}{\bar{t}} - \frac{f(\bar{y}+\bar{t}d) - f(\bar{y})}{\bar{t}}\Big| \leq \varepsilon L\|d\| + \varepsilon. \tag{5.7}$$

From (5.6), (5.5) and (5.7), it follows that

$$\begin{aligned}
\bar{f}^\circ(x;d) &\leq f^\circ(x;d) + \frac{f(y(\bar{t})) - f(\bar{y})}{\bar{t}} - \sup_{\|z-x\|<\delta, 0<s<t} \frac{f(z+sd) - f(z)}{s} + 2\varepsilon \\
&\leq f^\circ(x;d) + \frac{f(y(\bar{t})) - f(\bar{y})}{\bar{t}} - \frac{f(\bar{y}+\bar{t}d) - f(\bar{y})}{\bar{t}} + 2\varepsilon \\
&\leq f^\circ(x;d) + (3 + L\|d\|)\varepsilon.
\end{aligned}$$

Since ε is arbitrary, we conclude $\bar{f}^\circ(x;d) \leq f^\circ(x;d)$.

If (ii) does not hold, there must exist a vector η such that $\eta \in \partial \bar{f}(x)$ but $\eta \notin P_x \partial f(x)$. By Proposition 2.1.2 in Clarke (2003), $P_x \partial f(x) \subset T_x\mathscr{M}$ is a closed convex set. Note that $\partial \bar{f}(x) \subset T_x\mathscr{M}$. By the Convex Separation theorem, there exists $d \in T_x\mathscr{M}$ such that

$$\langle \eta, d\rangle > \sup_{\xi \in P_x \partial f(x)} \langle \xi, d\rangle. \tag{5.8}$$

Since P_x is a symmetric matrix and $P_x d = d$, we have

$$\sup_{\xi \in P_x \partial f(x)} \langle \xi, d\rangle = \sup_{\zeta \in \partial f(x)} \langle P_x\zeta, d\rangle = \sup_{\zeta \in \partial f(x)} \langle \zeta, d\rangle = f^\circ(x;d). \tag{5.9}$$

By (5.8) and (5.9), $\bar{f}^\circ(x;d) \geq \langle \eta, d\rangle > f^\circ(x;d)$, which contradicts (i).

(iii). From the proof of (ii), it suffices to prove that $\bar{f}^\circ(x;d) = f^\circ(x;d)$. Let (U,φ) a chart around x and let $x(t) = \varphi^{-1}(\varphi(x) + t\mathbf{D}\varphi(x)(d))$. Since φ is a diffeomorphism and $\mathbf{D}\varphi(x) = Id_{T_x\mathscr{M}}$, we have

$$x(t) = x + t\mathbf{D}\varphi^{-1}(\varphi(x))(\mathbf{D}\varphi(x)(d)) + O(t^2) = x + td + O(t^2). \tag{5.10}$$

Assume that the Lipschitz constant of f at x is L. Fix $\varepsilon > 0$. By (5.10), for sufficiently small t, we have

$$\Big|\frac{f(x+td) - f(x)}{t} - \frac{f(x(t)) - f(x)}{t}\Big| \leq \varepsilon L. \tag{5.11}$$

Since by assumption $f'(x;d)$ exists, we can also assume that for all sufficiently small $t > 0$,

$$f'(x;d) - \frac{f(x+td) - f(x)}{t} \leq \varepsilon.$$

Recall that $U_\delta := \{y \in \mathscr{M} : d(x,y) < \delta\}$. By (3.1) and $\bar{f} = f|_{\mathscr{M}}$, there exist $\delta > 0$ and $t_0 > 0$ such that

$$\sup_{y \in U_\delta, 0<t<t_0} \frac{f \circ \varphi^{-1}(\varphi(y) + t\mathbf{D}\varphi(x)(d)) - f(y)}{t} - \bar{f}^\circ(x;d) \leq \varepsilon. \tag{5.12}$$

From (5.10), (5.11) and (5.12), it follows that

$$f'(x;d) \leq \bar{f}^{\circ}(x;d) + (2+L)\varepsilon.$$

Since ε is arbitrary, we have $f'(x;d) \leq \bar{f}^{\circ}(x;d)$. Since f is regular at x along d, we have

$$f'(x;d) = f^{\circ}(x;d) \geq \bar{f}^{\circ}(x;d) \geq f'(x;d).$$

Thus $\bar{f}^{\circ}(x;d) = f^{\circ}(x;d)$ and the proof is complete.

According to Proposition 2.3.6 in Clarke (1983), we know that there are a variety of types of functions satisfying the conditions of Definition 5.1. In particular, we have the following lemma.

Lemma 5.1 *Let $f = f_1 + f_2$ be a function on $\mathbb{R}^m$, where f_1 is convex and f_2 is continuously differentiable. Then f is regular along $T_x\mathscr{M}$, where $x \in \mathscr{M}$ is arbitrary.*

In the remainder, we will discussion the connection between our optimization conditions on Riemannian manifold with the traditional ones by considering the more specific situation when

$$\mathscr{M} = \{x \in \mathbb{R}^m \mid \Phi(x) = 0\},$$

where $\Phi : \mathbb{R}^m \to \mathbb{R}^q$ $(q = m-n)$ is a smooth mapping with $\mathbf{D}\Phi(x)$ of full row rank for all $x \in \mathscr{M}$. For these problems, we can incorporate the particular manifold $\mathscr{M}$ into our original problem (1.2), and simply state it as the following traditional nonlinear programming problem:

$$\begin{array}{lll} \min & f(x) & \\ s.t. & c_i(x) = 0, & i \in \mathscr{E} = \{1, \cdots, l\} \\ & c_i(x) \geq 0, & i \in \mathscr{I} = \{l+1, \cdots, r\} \\ & \Phi(x) = 0. & \end{array} \tag{5.13}$$

When the problem (1.2) is put in this way, one then can directly apply the traditional optimality conditions without referring to the underlying manifold structure of $\mathscr{M}$. Then, we will describe the connection between our established optimality conditions in Section 4 with the traditional conditions, and show that our optimality conditions can be derived directly from the traditional optimality conditions. We believe that this connection could also be helpful for designing efficient numerical algorithms.

First, if we simply regard (5.13) as the traditional nonlinear programming, then the Lagrangian function is

$$L(x, \lambda, \mu) = f(x) - \sum_{i \in \mathscr{E} \cup \mathscr{I}} \lambda_i c_i(x) - \sum_{i=1}^{q} \mu_i \Phi_i(x),$$

and the traditional *linear independent constraint qualification* (LICQ)(see Definition 12.4 in Nocedal and Wright (2006)) at x is as follows:

$$\{\nabla c_i(x), i \in \mathscr{A}(x)\} \cup \{\nabla \Phi_j(x), 1 \leq j \leq q\} \text{ is linear independent on } \mathbb{R}^m. \tag{5.14}$$

On the other hand, under the condition that $\mathbf{D}\Phi(x)$ of full row rank for all $x \in \mathscr{M}$, it is well known that (see Section 3.5.7 in Absil et al. (2008))

$$T_x\mathscr{M} = \mathrm{Ker}(\mathbf{D}\Phi(x)), \quad \forall x \in \mathscr{M}.$$

Moreover, since $\mathbf{D}\Phi(x)$ has full row rank, the matrix

$$P_x := I_m - \mathbf{D}\Phi(x)^\top(\mathbf{D}\Phi(x)\mathbf{D}\Phi(x)^\top)^{-1}\mathbf{D}\Phi(x)$$

is the orthogonal projection onto $T_x\mathscr{M}$. If f is differentiable, using P_x, we have the relation

$$\mathrm{grad}_x\mathscr{L}(x,\lambda) = P_x\nabla_x L(x,\lambda,\mu),$$

where the function $\mathscr{L}(x,\lambda)$ is defined in (4.7). If f is convex, then by Theorem 5.1 and Lemma 5.1, we have

$$\partial_x\mathscr{L}(x,\lambda) = P_x\partial_x L(x,\lambda,\mu). \tag{5.15}$$

To show the connection between our optimality conditions in Section 4 with the traditional conditions, we establish the following three key lemmas, from which the equivalence of the two types of optimality conditions becomes evident.

Lemma 5.2 *For all $x \in \Omega$, the LICQ condition* (4.3) *holds if and only if the LICQ condition* (5.14) *holds.*

Proof Sufficiency: Assume that (4.3) holds. If (5.14) is not true, then there exist $\lambda_i,\ i \in \mathscr{E}\cup\mathscr{I}$, and $\mu_j,\ 1 \le j \le q$, such that

$$\sum_{i\in\mathscr{E}\cup\mathscr{I}} |\lambda_i| + \sum_{j=1}^{q} |\mu_j| > 0 \tag{5.16}$$

and

$$\sum_{i\in\mathscr{E}\cup\mathscr{I}} \lambda_i\nabla c_i(x) + \sum_{j=1}^{q} \mu_j\nabla\Phi_j(x) = 0. \tag{5.17}$$

Since $P_x\nabla\Phi_j(x) = 0$, (5.17) and (5.1) imply that

$$\sum_{i\in\mathscr{E}\cup\mathscr{I}} \lambda_i \mathrm{grad}\, c_i(x) = \sum_{i\in\mathscr{E}\cup\mathscr{I}} \lambda_i P_x\nabla c_i(x) = 0. \tag{5.18}$$

From (4.3), it follows that $\lambda_i = 0,\ \forall i \in \mathscr{E}\cup\mathscr{I}$, which together with (5.17) imply $\sum_{j=1}^{q}\mu_j\nabla\Phi_j(x) = 0$. This and the assumption that $\mathbf{D}\Phi(x)$ has full row rank imply that $\mu_j = 0,\ \forall 1 \le j \le q$, a contradiction to (5.16).

Necessity: If (4.3) does not hold, then there exist $\lambda_i,\ i \in \mathscr{E}\cup\mathscr{I}$, not all zero, such that (5.18) holds. Since P_x is the orthogonal projection onto $\mathrm{Ker}(\mathbf{D}\Phi(x))$, by $(\mathrm{Ker}(\mathbf{D}\Phi(x)))^\perp = \mathrm{Range}(\mathbf{D}\Phi(x)^\top)$, there exists a vector $\mu \in \mathbb{R}^q$ such that

$$\sum_{i\in\mathscr{E}\cup\mathscr{I}} \lambda_i\nabla c_i(x) = (\mathbf{D}\Phi(x))^\top\mu = \sum_{j=1}^{q} \mu_j\nabla\Phi_j(x).$$

Then (5.14) is not true and the proof is complete.

Related with the traditional nonlinear programming problem (5.13), the set of linearized feasible directions $\mathrm{F}(x)$ (see Definition 12.3 in Nocedal and Wright (2006)) is given by

$$\mathrm{F}(x) = \left\{ d \in \mathbb{R}^m \;\middle|\; \begin{array}{ll} \langle \nabla \Phi_i(x), d\rangle = 0, & \text{for all } 1 \le i \le q, \\ \langle \nabla c_i(x), d\rangle = 0, & \text{for all } i \in \mathscr{E}, \\ \langle \nabla c_i(x), d\rangle \ge 0, & \text{for all } i \in \mathscr{A}(x) \cap \mathscr{I} \end{array} \right\},$$

which is shown to be identical to $\mathscr{F}(x)$ defined by (4.5).

Lemma 5.3 *For all $x \in \mathscr{M}$, we have $\mathscr{F}(x) = \mathrm{F}(x)$.*

Proof Note that $\langle \nabla \Phi_i(x), d\rangle = 0$ for all $1 \le i \le q$ is equivalent to $d \in T_x\mathscr{M}$. This together with the fact

$$\langle \nabla c_j(x), d\rangle = \langle \nabla c_j(x), P_x d\rangle = \langle P_x \nabla c_j(x), d\rangle = \langle \mathrm{grad}\, c_j(x), d\rangle, \quad \forall d \in T_x\mathscr{M},$$

implies that $\mathscr{F}(x) = \mathrm{F}(x)$.

Recall that $\mathrm{N}_\Omega(x)$ is the normal cone to Ω at x in the traditional sence. If (4.3) holds, by Lemma 5.2, the LICQ condition (5.14) holds also, which together with Lemma 12.9 in Nocedal and Wright (2006) implies that

$$\mathrm{N}_\Omega(x) = \{ \sum_{i \in \mathscr{A}(x)} \lambda_i \nabla c_i(x) + \sum_{j=1}^{q} \mu_j \nabla \Phi_j(x), \quad \lambda_i \le 0 \text{ for } i \in \mathscr{A}(x) \cap \mathscr{I} \}. \tag{5.19}$$

Using $P_x \nabla \Phi_j(x) = 0$ again, we have the following result.

Lemma 5.4 *Let Ω be the feasible set of* (5.13) *and $x \in \Omega$. Suppose that the LICQ condition* (4.3) *is satisfied. Then*

$$\mathscr{N}_\Omega(x) = P_x \mathrm{N}_\Omega(x),$$

where $\mathscr{N}_\Omega(x)$ is defined by (4.6).

Equipped with the previous results and techniques, we are able to derive the optimality conditions in the preceding sections.

Consider the optimization problem (1.2), with $f = f_1 + f_2$, where f_1 is convex and f_2 is continuously differentiable, and the LICQ condition (4.3) holds. Then Propostion 3.3 follows from Corollary 2.4.3 in Clarke (1983), Theorem 5.1, Lemmas 5.1 and 5.4.

The first-order optimality condition in Theorem 4.1 directly follows from the traditional KKT conditions (see Corollary 2.4.3 in Clarke (1983)) for problem (5.13):

$$\begin{aligned} 0 &\in \partial_x L(x, \lambda, \mu), \\ \Phi(x) = 0, \quad c_i(x) &= 0, \quad \text{for all } i \in \mathscr{E}, \\ c_i(x) \ge 0, \quad \lambda_i \ge 0, \quad \lambda_i c_i(x) &= 0, \quad \text{for all } i \in \mathscr{I}, \end{aligned} \tag{5.20}$$

where $\lambda \in \mathbb{R}^r$ and $\mu \in \mathbb{R}^q$. In fact, if the LICQ condition (4.3) holds, by Lemma 5.2, it follows that the LICQ condition (5.14) is true, and so $\mathrm{N}_\Omega(x^*)$ is given by (5.19). By Corollary 2.4.3 in Clarke (1983),

there exists (x^*,λ^*,μ^*) satisfy the KKT condition (5.20), which, together with (5.15), shows that (x^*,λ^*) satisfy the KKT conditions given in Theorem 4.1.

Now assume that f and c_i for $i\in\mathscr{E}\cup\mathscr{I}$ are twice differentiable functions. For the second-order optimality conditions, we suppose that (x^*,λ^*,μ^*) satisfy the KKT condition (5.20). Then the traditional *critical cone* $\mathrm{C}(x^*,\lambda^*,\mu^*)$ of problem (5.13) is defined by (see Section 12.5 in Nocedal and Wright (2006))

$$\mathrm{C}(x^*,\lambda^*,\mu^*)=\{w\in F(x^*)\;\Big|\;\begin{array}{ll}\langle\nabla\Phi_i(x^*),d\rangle=0, & \text{for all } 1\le i\le q,\\ \langle\nabla c_i(x^*),w\rangle=0, & \text{all } i\in\mathscr{A}(x^*)\cap\mathscr{I} \text{ with } \lambda_i^*>0\end{array}\}.$$

Similar to the proof of Lemma 5.3, it is easy to see that

$$\mathscr{C}(x^*,\lambda^*)=\mathrm{C}(x^*,\lambda^*,\mu^*), \tag{5.21}$$

where $\mathscr{C}(x^*,\lambda^*)$ is given in (4.9). Moreover, given $x\in\mathscr{M}$ and $\lambda\in\mathbb{R}^r$, if the LICQ condition (5.14) holds at x, then the system $\nabla_x L(x,\lambda,\mu)=0$ in (5.20) admits a unique least-squares solution

$$\mu(x,\lambda)=-\mathbf{D}\Phi(x)^\top(\mathbf{D}\Phi(x)\mathbf{D}\Phi(x)^\top)^{-1}\mathbf{D}\Phi(x)(\nabla f(x)-\sum_{i\in\mathscr{E}\cup\mathscr{I}}\lambda_i\nabla c_i(x)). \tag{5.22}$$

Note that the Hessian of the restriction $h|_{\mathscr{M}}$ at $x\in\mathscr{M}$ can be calculated via (see Proposition 5.3.2 in Absil et al. (2008))

$$\begin{aligned}\mathrm{Hess}h|_{\mathscr{M}}(x)&=P_x\mathbf{D}(\mathrm{grad}h|_{\mathscr{M}})(x)\\&=P_x\mathbf{D}(P_x\nabla h)(x).\end{aligned} \tag{5.23}$$

For simplifying the following presentation, when the meaning is clear from the context, we will also use h to denote the restriction $h|_{\mathscr{M}}$ of h on $\mathscr{M}$. Additionally, we will use $\mathrm{grad}h(x)$ and $\nabla h(x)$ to distinguish the gradients of $h|_{\mathscr{M}}$ and h at x, respectively, and similarly, use $\mathrm{Hess}h(x)$ and $\nabla^2 h(x)$ to distinguish the Hessians of $h|_{\mathscr{M}}$ and h, respectively.

For any $x\in\Omega$, $\lambda\in\mathbb{R}^r$ and $w\in T_x\mathscr{M}$, by the technique used in Absil et al. (2009) and (5.23), we have that

$$\begin{aligned}&\mathrm{Hess}_x\mathscr{L}(x,\lambda)[w]\\&=P_x\big[\mathbf{D}_x(P_x\nabla\mathscr{L}(x,\lambda))\big][w] && \text{(by (5.23))}\\&=P_x\big[\mathbf{D}_x(\nabla\mathscr{L}(x,\lambda)-\mathbf{D}\Phi(x)^\top(\mathbf{D}\Phi(x)\mathbf{D}\Phi(x)^\top)^{-1}\mathbf{D}\Phi(x)\nabla\mathscr{L}(x,\lambda))\big][w]\\&=P_x\big[\mathbf{D}_x(\nabla\mathscr{L}(x,\lambda)-\mathbf{D}\Phi(x)^\top\mu(x,\lambda))\big][w] && \text{(by (5.22))}\\&=P_x\big[\nabla^2_{xx}\mathscr{L}(x,\lambda)-\sum_{i=1}^q\mu_i(x,\lambda)\nabla^2_{xx}\Phi_i(x)-\mathbf{D}\Phi(x)^\top\mathbf{D}_x\mu(x,\lambda)\big][w]\\&=P_x\nabla^2_{xx}L(x,\lambda,\mu(x,\lambda))P_x[w],\end{aligned}$$

where the last equality follows due to $P_x\mathbf{D}\Phi(x)^\top=0$ and $P_xw=w$. Thus, we have

$$\mathrm{Hess}_x\mathscr{L}(x,\lambda)=P_x\nabla^2_{xx}L(x,\lambda,\mu(x,\lambda))P_x. \tag{5.24}$$

Therefore, if x^* is a local solution of (1.2) with Lagrange multipliers (λ^*, μ^*) satisfying (5.20), from (5.24), it follows that

$$\mathrm{Hess}_{x^*}\mathscr{L}(x^*, \lambda^*) = P_{x^*}\nabla^2_{xx}L(x^*, \lambda^*, \mu^*)P_{x^*}.$$

As a result, with the aid of (5.21), and $P_{x^*}w = w$ for any $w \in \mathrm{C}(x^*, \lambda^*, \mu^*)$, Theorems 4.2 and 4.3 can then be obtained from the traditional second-order optimality conditions Theorem 12.5 and Theorem 12.6 of Nocedal and Wright (2006), respectively.

6 Conclusion

In this paper, we formulated the general nonlinear programming that is built upon a general Riemannian manifold, and established its optimality conditions. We showed that, in the language of differential geometry, these optimality conditions coincide with the traditional conditions for the nonlinear programming. This result, on the one hand, sheds some lights on the underlying optimization problem, and on the other hand, lays the ground for further generalizing other classical optimization methods to the manifold-based nonlinear programming.

References

1. Absil P-A, Baker CG, Gallivan KA (2007) Trust-region methods on Riemannian manifolds. Found Comput Math 7:303–330
2. Absil P-A, Mahony R, Sepulchre R (2008) Optimization algorithms on matrix manifolds. Princeton University Press, Princeton
3. Absil P-A, Mahony R, Sepulchre R, Van Dooren P (2002) A Grassmann–Rayleigh quotient iteration for computing invariant subspaces. SIAM Rev 44:57–73
4. Absil P-A, Trumpf J, Mahony R, Andrews B (2009) All roads lead to Newton: feasible second-order methods for equality-constrained optimization. Technical report, Department d'ingenierie mathematique, UCLouvain, Belgium
5. Adler RL, Dedieu J-P, Margulies JY, Martens M, Shub M (2002) Newton's method on Riemannian manifolds and a geometric model for the human spine. IMA J Numer Anal 22:359–390
6. Azagra D, Ferrera J, López-Mesas F (2005) Nonsmooth analysis and Hamilton-Jacobi equations on Riemannian manifolds. J Funct Anal 220:304–361
7. Clarke FH (1983) Optimization and Nonsmooth Analysis. Wiley, New York
8. Edelman A, Arias TA, Smith ST (1998) The geometry of algorithms with orthogonality constraints. SIAM J Matrix Anal Appl 20:303–353
9. Ferreira OP (2006) Proximal subgradient and a characterization of Lipschitz function on Riemannian manifolds. J Math Anal Appl 313:587–597
10. Ferreira OP (2008) Dini derivative and a characterization for Lipschitz and convex functions on Riemannian manifolds. Nonlinear Anal 68:1517–1528
11. Gabay D (1982) Minimizing a differentiable function over a differential manifold. J Optim Theory Appl 37:177–219
12. Hosseini S, Pouryayevali MR (2011) Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds. Nonlinear Anal 74:3884–3895
13. Ledyaev YS, Zhu QJ (2007) Nonsmooth analysis on smooth manifolds. Trans Amer Math Soc 359:3687–3732
14. Lee JM (2003) Introduction to smooth manifolds. Springer-Verlag, New York
15. Lu Z, Zhang Y (2012) An augmented Lagrangian approach for sparse principal component analysis. Math Program 135:149C193

16. Klingenberg W (1982) Riemannian geometry. Berlin-New York
17. Mahony RE (1996) The constrained Newton method on a Lie group and the symmetric eigenvalue problem. Linear Algebra Appl 248:67–89
18. Motreanu D, Pavel NH (1982) Quasitangent vectors in flow-invariance and optimization problems on Banach manifolds. J Math Anal Appl 88:116–132
19. Nocedal J, Wright SJ (2006) Numerical optimization. Springer Verlag, New York
20. Wen ZW, Yin WT (2010) A feasible method for optimization with orthogonal constraints. Rice University CAAM Technical Report TR10-26
21. Udriste C (1988) Kuhn-Tucker theorem on Riemannian manifolds. Topics in differential geometry, Vol. I, II (Debrecen, 1984), 1247–1259, North-Holland, Amsterdam
22. Yang Y (2007) Globally convergent optimization algorithms on Riemannian manifolds. J Optim Theory Appl 132:245–265
23. Zhang L-H (2010) Riemannian Newton method for the multivariate eigenvalue problem. SIAM J Matrix Anal Appl 31:2972–2996