J. Nonlinear Var. Anal. 10 (2026), No. 3, pp. 617-633
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.10.2026.3.06

ON THE CONVERGENCE AND PROPERTIES OF A PROXIMAL-GRADIENT METHOD ON HADAMARD MANIFOLDS

GLAYDSTON C. BENTO[1,*], CLAUDEMIR R. SANTIAGO[1,2]

[1] *Institute of Mathematics and Statistics, Federal University of Goiás, 74.690-900-Goiânia, Brazil*
[2] *CCHS, Federal Institute of Sergipe, 49.400-000-Lagarto, Brazil*

Abstract. In this paper, we address composite optimization problems on Hadamard manifolds, where the objective function is given by the sum of a smooth term (not necessarily convex) and a convex term (not necessarily differentiable). To solve this problem, we develop a proximal gradient method defined directly on the manifold, employing a strategy that enforces monotonicity of the objective function values along the generated sequence. We investigate its convergence properties without imposing the Lipschitz continuity assumption on the gradient of the smooth component.

Keywords. Composite optimization; Hadamard manifolds; Riemannian optimization; Riemannian proximal-gradient method.

1. INTRODUCTION

In this paper, we are interested in the following optimization problem

$$\min_{x\in\mathscr{M}} F(x) := f(x) + g(x), \tag{1.1}$$

where $f:\mathscr{M}\to\mathbb{R}$ is differentiable and has continuous gradient, $g:\mathscr{M}\to\mathbb{R}$ is convex, not necessarily differentiable, and $\mathscr{M}$ is a finite dimensional Hadamard manifold. To the best of our knowledge, one of the earliest proposals of a proximal gradient-type method in the Euclidean setting was introduced in [17] as a generalization of the proximal point algorithm, designed to address the non-convexity of the objective function F, which is composed of a continuously differentiable component f and a convex component g, by linearizing f at each iteration. More precisely, the authors in [17] considered, from an initial point x_0, a sequence of points x_k generated according to the following scheme:

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\langle\nabla f(x_k), x-x_k\rangle + \frac{1}{2\alpha_k}\|x_k-x\|^2 + g(x)\right\}, \tag{1.2}$$

where $\{\alpha_k\}$ is sequence of positive numbers. It can be observed that (1.2) is equivalently to the following formulation

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\frac{1}{2\alpha_k}\|x-(x_k-\alpha_k\nabla f(x_k))\|^2 + g(x)\right\}, \tag{1.3}$$

*Corresponding author.
E-mail address: glaydston@ufg.br (G.C. Bento), claudemir.santiago@ifs.edu.br (R.S. Claudemir).
Received 28 September 2025; Accepted 15 December 2025; Published online 1 April 2026.
©2026 Journal of Nonlinear and Variational Analysis

which is commonly discussed in the literature (see [3, 4]) and often referred to as the Iterative Shrinkage/Thresholding Algorithm (ISTA).

In recent years, the extension and analyses of optimization algorithms within the context of Riemannian manifolds have attracted the attention from many researchers; see, e.g., [1, 2, 5, 7, 12, 15, 26] and the references therein.

The proximal gradient method can be formulated in various ways, each with properties that present distinct challenges, particularly in their analysis in the context of Hadamard manifolds. In [11], a proximal gradient method based on the formulation given in (1.2) was proposed for solving problems in which the objective function has a composite structure: f is a smooth (possibly non-convex) function with a continuous Lipschitz gradient, and g is convex, not necessarily smooth, defined on the Stiefel manifold. Specifically, they proposed solving the following subproblem to determine a descent direction v_k at the k-th iteration:

$$v_k := \operatorname*{argmin}_{v \in T_{x_k}\mathscr{M}} \left\{ \langle \nabla f(x_k), v\rangle + \frac{1}{2\alpha_k}\|v\|_E^2 + g(x_k+v) \right\}. \tag{1.4}$$

Note that the Stiefel manifold $\mathscr{M}$ is embedded in $\mathbb{R}^{n\times r}$. Hence, a point x on the manifold M and an element v on its corresponding tangent space $T_x\mathscr{M}$ are identified as elements in $\mathbb{R}^{n\times r}$ ensuring a well-defined meaning for the sum $x+v$ thereby giving proper interpretation to the term $g(x+v)$ in the formulation above. Moreover, in the approach proposed by the authors, the Riemannian gradient is not computed; instead, only the Euclidean gradient is evaluated, and the sequence x_k is bounded due to the compactness of the manifold, and a linear search α_k is performed satisfying an Armijo-type condition. They proved that every limit point of the sequence x_k generated by the algorithm is a stationary point of the problem, and that the algorithm has sublinear iteration complexity to find an ε-stationary point. In [19], the authors introduced and analyzed a Riemannian proximal gradient method on general Riemannian manifolds. Instead of the proximal gradient formulation in (1.4) previously considered in [11], the authors employed a distinct Riemannian proximal mapping in which the term $x+v$ is replaced by a retraction $R_x(v)$, and an inexact criterion was considered in order to solve the proximal subproblem. Specifically, the authors first defined the function

$$\ell_{x_k}(v) := \langle \operatorname{grad} f(x_k), v\rangle_{x_k} + \frac{L}{2}\langle v, v\rangle_{x_k} + g(R_{x_k}(v)),$$

and proposed to compute a direction $v_k \in T_{x_k}\mathscr{M}$ such that

$$v_k \text{ is a stationary point of } \ell_{x_k}(v) \text{ on } T_{x_k}\mathscr{M} \text{ satisfying } \ell_{x_k}(v_k) \le \ell_{x_k}(0). \tag{1.5}$$

Under the assumption of global Lipschitz continuity of $\operatorname{grad} f$, the authors proved that any accumulation point of the sequence generated by their method is a stationary point. Subsequently, [22] analyzed a proximal gradient method similar to the one in [19], proving asymptotic convergence and establishing, in the convex case, a sublinear convergence rate in terms of functional values. In [14], Feng et al. presented different versions of the proximal gradient mapping, which correspond to the respective extensions to Riemannian manifolds of (1.2) and (1.3) respectively, i.e.,

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathscr{M}} \left\{ \left\langle \operatorname{grad} f(x_k), \exp_{x_k}^{-1} x \right\rangle + \frac{1}{2\alpha}\left\|\exp_{x_k}^{-1} x\right\|^2 + g(x) \right\}$$

and

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathscr{M}} \left\{ g(x) + \frac{1}{2\alpha} \left\| \exp_{\hat{x}_k}^{-1} x \right\|^2 \right\},$$

where $\hat{x}_k = \exp_{x_k}(-\alpha \operatorname{grad} f(x_k))$. In this context, they replaced the proximal gradient mapping used in [19] with an exact version (see Definition 3.1) and proposed a proximal gradient method that solves a subproblem directly on the manifold. This approach is a generalization of the proximal gradient algorithm presented in [21], now adapted to Hadamard manifolds. Partial convergence of the proposed algorithm was demonstrated, and the techniques employed take advantage of the special geometric structure of these manifolds, still under the assumption of global Lipschitz continuity of the gradient of f and a constant step size. The results show that the algorithm is effectively convergent for minimizing problems of the form (1.1), even when the objective function is nonconvex and nonsmooth. Recently, in [9], the authors considered a intrinsic variant of the Riemannian proximal gradient method introduced in [19], in which, broadly speaking, the auxiliary function (1.5) defined on the tangent bundle of the manifold is replaced by an auxiliary function defined directly on the manifold, thus ensuring a completely intrinsically approach. The main results are established under the local Lipschitz continuity of one of the component functions and the bounded sectional curvature of the manifold, while not requiring convexity of the functions f and g.

In this paper, we consider a Riemannian proximal gradient method for solving Problem (1.1) and analyze its convergence on Hadamard manifold without assuming local Lipschitz continuity of the grad f. A backtracking scheme is considered in order to compute a step size and generate the next iterate. This scheme is important for problems in which the Lipschitz constant, denoted by L, of grad f does not exist or is hard to compute. Indeed, the use of such backtracking scheme may be convenient even in situations in which the Lipschitz constant L does exist but it is very large. Hence, the use of such a scheme usually provides a larger step size than the usual $1/L$ constant step size.

This paper is organized as follows: Section 2 contains notation, definitions and basic results on manifolds, as well as some results of convex analysis and differentiability on Hadamard manifolds. Section 3 establishes the Riemannian proximal gradient method and presents basic properties. Section 4 is devoted to the convergence analysis of the method. Section 5 contains a conclusion.

2. Preliminaries

2.1. **Notations.** In this section, we review some standard definitions and results that serve as support throughout the paper and that can be found in several classical texts (see, e.g., [10, 13, 25]).

Let $\mathscr{M}$ be a Hadamard manifold, i.e,, a complete simply connected Riemannian manifold of nonpositive sectional curvature, $T_x\mathscr{M}$ the tangent space of $\mathscr{M}$ at $x \in \mathscr{M}$ and $T\mathscr{M} = \bigcup_{x \in \mathscr{M}} T_x\mathscr{M}$ the tangent bundle of $\mathscr{M}$. The parallel transport $\mathscr{P}_{\gamma(a),\gamma(b)} : T_{\gamma(a)}\mathscr{M} \to T_{\gamma(b)}\mathscr{M}$ on the tangent bundle T$\mathscr{M}$ along $\gamma : [a,b] \to \mathbb{R}$ with respect to ∇ is defined by

$$\mathscr{P}_{\gamma(a),\gamma(b)} v = X(\gamma(b)), \ \forall a,b \in \mathbb{R} \text{ and } v \in T_{\gamma(a)}\mathscr{M},$$

where X is the unique vector field such that

$$\nabla_{\gamma'(t)} X = 0, \forall t \in [a,b] \text{ and } X(\gamma(a)) = v.$$

Proposition 2.1. *The parallel transport operator $\mathscr{P}_{\gamma(a),\gamma(b)}$ is linear. Also, for any $a,\tilde{b},b\in\mathbb{R}$, $\mathscr{P}_{\gamma(\tilde{b}),\gamma(b)}\circ\mathscr{P}_{\gamma(a),\gamma(\tilde{b})}=\mathscr{P}_{\gamma(a),\gamma(b)}$ and $\mathscr{P}_{\gamma(a),\gamma(a)}$ is the identity. In particular, the inverse of $\mathscr{P}_{\gamma(a),\gamma(b)}$ is $\mathscr{P}^{-1}_{\gamma(b),\gamma(a)}$. If $\mathscr{M}$ is a Riemannian manifold and ∇ is compatible with the Riemannian metric, then parallel transport is an isometry, that is,*

$$\left\langle \mathscr{P}_{\gamma(a),\gamma(b)}u,\mathscr{P}_{\gamma(a),\gamma(b)}v\right\rangle_{\gamma(b)}=\langle u,v\rangle_{\gamma(a)}\,,\ \forall a,b\in\mathbb{R}\text{ and }v\in T_{\gamma(a)}\mathscr{M}\,.$$

Stated differently, the adjoint and the inverse of $\mathscr{P}_{\gamma(a),\gamma(b)}$ coincide.

Since it is an isometry between the respective tangent spaces, preserves norm and, roughly speaking, a "direction", analogous to translation in $\mathbb{R}^n$. A vector tangent to a geodesic γ remains tangent if transported parallel along it. If $\gamma:[a,b]\to\mathscr{M}$ is a piecewise differentiable curve joining $x=\gamma(a)$ to $y=\gamma(b)$ in $\mathscr{M}$, we define the length of the curve γ as $L(\gamma)=\int_x^y\|\gamma'(t)\|\,dt$. The minimal length of all such curves joining x to y is called the Riemannian distance, and it is denoted by $d(x,y)$.

An exponential map at $x\in\mathscr{M}$ is a mapping defined by $\exp_x:T_x\mathscr{M}$ to $\mathscr{M}$ such that $\exp_x tv=\gamma_v(t;x)$; $\forall\, v\in T_x\mathscr{M}$ and $t\in\mathbb{R}$, where $\gamma_v(\cdot;x)$ is the geodesic starting from x with velocity v, i.e., $\gamma_v(0;x)=x$, $y=\gamma_v(1;x)$ and $\gamma_v'(0;x)=v$. It follows that $\exp_x 0=\gamma_v(0;x)=x$ where 0 is the zero tangent vector. Recall that the exponential map $\exp_x(\cdot)$ is differentiable on $T_x\mathscr{M}$ for any $x\in\mathscr{M}$. Therefore, by the Inverse Mapping Theorem, we define the inverse exponential map $\exp_x^{-1}:\mathscr{M}\to T_x\mathscr{M}$, and in addition, for any $x,y\in\mathscr{M}$ we have $d(x,y)=\left\|\exp_x^{-1}y\right\|=\left\|\exp_y^{-1}x\right\|$. Since $\mathscr{M}$ is a Hadamard manifold, we have for any $t\in[0,1]$ that the geodesic joining x to y can be defined as $\gamma(t)=\exp_x\left(t\exp_x^{-1}y\right)$.

Example 2.1. Take $x\in\mathscr{M}$ and let $\exp_x^{-1}:\mathscr{M}\to T_x\mathscr{M}$ be the inverse of the exponential map; the map $d^2(\cdot,x):\mathscr{M}\to\mathbb{R}$ is strongly convex, C^∞ and $\operatorname{grad}\frac{1}{2}d^2(y,x)=-\exp_y^{-1}x$.

Theorem 2.1. *Let $\mathscr{M}$ be a complete Riemannian manifold with nonpositive sectional curvature. Then, for any point $x\in\mathscr{M}$, the exponential map is a covering map. Suppose further that $\mathscr{M}$ is simply connected. Then, for every $x\in\mathscr{M}$, the exponential map is a diffeomorphism. Moreover, for any two points $x,y\in\mathscr{M}$, there exists a unique normal geodesic joining x to y, which is minimal, i.e., it realizes the Riemannian distance between x and y.*

Proposition 2.2. *Let $\mathscr{M}$ be a Hadamard manifold and $\triangle y_1,y_2,y_3$ a geodesic triangle. Denote, for each $j=1,2,3\,(mod\,3)$, by $\gamma_j:[0,l_j]\to M$ geodesic joining y_j to y_{j+1}, set $d\left(y_j,y_{j+1}\right):=L(\gamma_j)$ and $\theta_j:=\angle(\gamma_j'(0),-\gamma_{j-1}'\left(l_{j-1}\right)$. Then, the Comparison Theorem for Triangles can be rewritten in terms of the distance and the exponential map as:*

$$d^2\left(y_{j+2},y_j\right)\geq d^2\left(y_j,y_{j+1}\right)+d^2\left(y_{j+1},y_{j+2}\right)-2\left\langle\exp_{y_{j+1}}^{-1}y_j,\exp_{y_{j+1}}^{-1}y_{j+2}\right\rangle;\tag{2.1}$$

$$d^2\left(y_j,y_{j+1}\right)\leq\left\langle\exp_{y_j}^{-1}y_{j+1},\exp_{y_j}^{-1}y_{j+2}\right\rangle+\left\langle\exp_{y_{j+1}}^{-1}y_j,\exp_{y_{j+1}}^{-1}y_{j+2}\right\rangle;\tag{2.2}$$

since $\left\langle\exp_{y_{j+1}}^{-1}y_j,\exp_{y_{j+1}}^{-1}y_{j+2}\right\rangle=d\left(y_j,y_{j+1}\right)\cdot d\left(y_{j+1},y_{j+2}\right)\cdot cos\left(\theta_{j+1}\right)$.

2.2. **Convexity and Directional Differentiability in Hadamard Manifolds.** In this subsection, we introduce some standard results of convex analysis and differentiability on Hadamard manifolds, which serve as support throughout the paper and can be found in several papers and classical texts, such as [6, 7, 8, 18, 27].

Given $c \in \mathbb{R}$, the sub-level set M^c is the set $\{x \in \mathscr{M} : F(x) \leq c\}$. In addition, given a scalar $\delta > 0$, the open and closed balls in M are denoted by $\mathscr{B}(x_0, \delta) = \{x \in \mathscr{M}; d(x_0, x) < \delta\}$ and $\bar{\mathscr{B}}(x_0, \delta) = \{x \in \mathscr{M}; d(x_0, x) \leq \delta\}$. A subset $\Omega \subseteq \mathscr{M}$ is said to be convex if, for any two points x and y in Ω, the geodesic joining x to y is contained in Ω.

Definition 2.1. A function $h : \mathscr{M} \to \mathbb{R}$ is said to be convex if for any $x, y \in \mathscr{M}$ and any geodesic γ such that $\gamma(0) = x$ and $\gamma(1) = y$, and $t \in [0,1]$, it holds that

$$h(\gamma(t)) \leq (1-t)h(x) + th(y), \tag{2.3}$$

and μ-strongly convex if it holds that

$$f(\gamma(t)) \leq (1-t)f(x) + tf(y) - \frac{\mu}{2}t(1-t)d^2(x,y).$$

The definition of convexity (2.3) is equivalent to the existence of a tangent vector $\eta_x \in T_x\mathscr{M}$ such that

$$h(y) \geq h(x) + \left\langle \eta_x, \exp_x^{-1} y \right\rangle_x, \qquad \forall x, y \in \mathscr{M}.$$

The vector η_x, as above, is a subgradient of h at x, and the set of all subgradients of h at x is the subdifferential of h at x, denoted by $\partial h(x)$. It is known that the subdiferential $\partial f(x)$ reduces to the gradient of h, denoted by $\operatorname{grad} h(x)$, when h is differentiable at x. Note that $\langle \cdot, \cdot \rangle_x$ denotes the inner product in the tangent space of x induced by the Riemannian metric. In the rest of the paper we will omit the sub-index x of the inner product when it is clear from the context.

A function $H : \mathscr{M} \to \mathbb{R}$ is said to be 1-*coercive* in $x \in \mathscr{M}$, if $\lim\limits_{d(x,y)\to\infty} \frac{H(y)}{d(x,y)} = \infty$.

Definition 2.2. Let $\gamma : I \to \mathscr{M}$ be a smooth curve such that $\gamma(0) = x$ and $\gamma'(0) = v$. The directional derivative of a real-valued function h, defined on a manifold $\mathscr{M}$, at a point x, is given by

$$dh(x)[v] = \lim_{t \to 0^+} \frac{h(\gamma(t)) - h(x)}{t}.$$

Furthermore, we define the Riemannian gradient $\operatorname{grad} h$ of a function differentiable $h : \mathscr{M} \to \mathbb{R}$ on a Riemannian manifold $\mathscr{M}$, at a point x, is the unique tangent vector at x, such that $\langle \operatorname{grad} h(x), v \rangle = dh(x)[v]$ for all $v \in T_x\mathscr{M}$, where $dh(x)$ denotes the differential of h at x.

Remark 2.1. It follows from [16, Theorem 2.3] that the set $\partial h(x)$ is nonempty, convex, and compact for each $x \in \mathscr{M}$.

Proposition 2.3. *Let $\{x_k\} \subset \mathscr{M}$ be a bounded sequence. If the sequence $\{v_k\}$ is such that $v_k \in \partial h(x_k)$ for each $k \in \mathbb{N}$, then $\{v_k\}$ is also bounded.*

Definition 2.3. A point x is critical to Problem 1.1 if the following inclusion holds:

$$-\operatorname{grad} f(x) \in \partial g(x).$$

In particular, if x is a local minimum, then x is a critical point of Problem 1.1.

3. A Riemannian Proximal-Gradient Method

In this section, we propose a proximal gradient algorithm on Hadamard manifolds, based on an adaptation of the algorithm in [20] (with minor modifications), together with the Riemannian proximal gradient mapping introduced in [14, Definition 5] to solve the subproblems directly on the manifold. We show that, if the algorithm terminates after a finite number of iterations, it finds a critical point of Problem (1.1). We also establish the well-definedness of the algorithm, ensuring that its inner linesearch procedure stops after a finite number of inner iterations whenever the current point is not a critical point of Problem (1.1). Finally, two additional technical lemmas are presented in this section.

Definition 3.1. Let $x \in \mathscr{M}$ and $\alpha > 0$. The proximal gradient mapping consists of first computing a gradient step $\hat{x} := \exp_x(-\alpha \operatorname{grad} f(x))$ followed by solving the proximal subproblem

$$\tilde{x} = prox_{\alpha g}(\hat{x}) = \operatorname*{arg\,min}_{x \in \mathscr{M}} \left\{ g(x) + \frac{1}{2\alpha} d^2(x, \hat{x}) \right\}. \tag{3.1}$$

Since g is a convex function and $d^2(\cdot, \hat{x})$ is strongly convex, then the objective function of the proximal subproblem (3.1) is also strongly convex, which ensures that it has a unique solution.

3.1. Riemannian proximal gradient method and its basic properties.

We start this subsection by stating the Riemannian proximal gradient method considered in this paper.

Algorithm 1(Riemannian Proximal Gradient Method):

Step 0. (Initialization) Let $x_0 \in \mathscr{M}$, $0 < \alpha_{\min} \leq \alpha_{\max} < \infty$, and $\tau, \sigma \in (0,1)$; set $k, i = 0$.

Step 1. (Initial stepsize) Choose $\alpha_k^0 \in [\alpha_{\min}, \alpha_{\max}]$;

Step 2. (Proximal Subproblem) compute $\tilde{x}_{k,i}$ such that:

$$0 \in \partial g(\tilde{x}_{k,i}) + \frac{1}{2\alpha_{k,i}} \operatorname{grad} d^2(\tilde{x}_{k,i}, \hat{x}_{k,i}), \tag{3.2}$$

where

$$\hat{x}_{k,i} := \exp_{x_k}\big(-\alpha_{k,i} \operatorname{grad} f(x_k)\big), \quad \text{and } \alpha_{k,i} := \tau^i \alpha_k^0;$$

Step 3. (Stopping Criterion) If $\tilde{x}_{k,i} = x_k$, then stop and output x_k;

Step 4. (Update rule) If the following decreasing condition holds

$$F(\tilde{x}_{k,i}) + \frac{\sigma}{2\alpha_{k,i}} d^2(x_k, \tilde{x}_{k,i}) \leq F(x_k), \tag{3.3}$$

then set $i_k := i$, $\alpha_k := \alpha_{k,i_k}$, $\hat{x}_k := \hat{x}_{k,i_k}$, $x_{k+1} := \tilde{x}_{k,i_k}$, update $k \leftarrow k+1$ and return to step 1; Otherwise, set $i = i+1$ and return to step 2.

We start by making some observations about Algorithm 1. First, note that it has two types of iterations: the outer iterations, denoted by the index k, and the inner iterations, denoted by the index i. Second, the proximal inclusion (3.2) is equivalent to the proximal subproblem (3.1) with $x = \tilde{x}_{k,i}$ and $\alpha = \alpha_{k,i}$, which involves first computing an intermediate point $\hat{x}_{k,i}$ via a gradient step, and then solving the corresponding proximal subproblem to obtain $\tilde{x}_{k,i}$. Moreover, note that, for every pair of outer and inner iterations (k,i), there exists a unique proximal solution $\tilde{x}_{k,i}$. Third, it will be shown in Proposition 3.1 that the stopping criterion holds if and only if the current iterate x_k is a critical solution to Problem (1.1) in the sense of Definition 2.3.

Fourth, Lemma 3.3 shows that if x_k is not a critical solution, then the algorithm is well-defined, i.e., inequality (3.3), which is used to update the next iterate, holds after a finite number of inner iterations. Fifth, at the beginning of each outer iteration k, the stepsize is restarted to an arbitrary value in the interval $[\alpha_{\min}, \alpha_{\max}]$ to control its magnitude. This procedure ensures that $\alpha_k \leq \alpha_{\max}$ for all $k \geq 0$. However, the subsequence $\alpha_{k,i}$ may eventually become sufficiently small, potentially satisfying $\alpha_{k,i} < \alpha_{\min}$, because the trial stepsize $\alpha_{k,i}$ decreases during the inner iterations. Our convergence analysis covers both cases, whether the sequence of stepsizes α_k converges to zero or not.

Remark 3.1. It is worth noting that, from the acceptance criterion (3.3) and the definition of x_{k+1}, it follows that the algorithm generates a sequence of decreasing functional values $\{F(x_k)\}$.

We begin by stating and proving a basic result guaranteeing that, if Algorithm 1 terminates, it finds a critical point of Problem (1.1).

Proposition 3.1. *Let $\{x_k\}$ be the sequence generated by Algorithm 1 and consider the direction $v_k := \exp_{x_k}^{-1} \tilde{x}_{k,i}$. Then, $v_k = 0$ if and only if x_k is a solution to Problem (1.1).*

Proof. Assume that $v_k = 0$. Then, we clearly have $\tilde{x}_{k,i} = x_k$. Hence, it follows from (3.2) with $i = i_k$ and the last relation in Example 2.1 that

$$0 \in \frac{1}{2\alpha_k} \operatorname{grad} d^2(x_k, \hat{x}_k) + \partial g(x_k) = -\frac{1}{\alpha_k} \exp_{x_k}^{-1} \hat{x}_k + \partial g(x_k),$$

which, in view of the definition of $\hat{x}_k$, yields

$$\frac{1}{\alpha_k} \exp_{x_k}^{-1} \left(\exp_{x_k} \left(-\alpha_k \operatorname{grad} f(x_k) \right) \right) \in \partial g(x_k).$$

Hence, we have $-\operatorname{grad} f(x_k) \in \partial g(x_k)$, implying, by definition 2.3, that x_k is a critical point of Problem (1.1).

Assume now that x_k is a solution of Problem (1.1). Thus, in view of the definition of $\hat{x}_k$, we have

$$-\operatorname{grad} f(x_k) = \frac{1}{\alpha_k} \exp_{x_k}^{-1} \hat{x}_k \in \partial g(x_k). \tag{3.4}$$

It follows by (3.2), the definition of $\hat{x}_k$, and the last relation in Example 2.1 that

$$-\operatorname{grad} \frac{1}{2\alpha_{k,i}} d^2\left(\tilde{x}_{k,i}, \hat{x}_{k,i}\right) = \frac{1}{\alpha_{k,i}} \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i} \in \partial g\left(\tilde{x}_{k,i}\right). \tag{3.5}$$

Using (3.4), (3.5), and the monotonicity of the subdifferential of g, we obtain

$$\left\langle \exp_{x_k}^{-1} \hat{x}_{k,i}, \exp_{x_k}^{-1} \tilde{x}_{k,i} \right\rangle + \left\langle \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i}, \exp_{\tilde{x}_{k,i}}^{-1} x_k \right\rangle \leq 0. \tag{3.6}$$

On the other hand, using inequality (2.2) with $y_j = x_k$, $y_{j+1} = \tilde{x}_{k,i}$ and $y_{j+2} = \hat{x}_{k,i}$, we have

$$\left\langle \exp_{x_k}^{-1} \hat{x}_{k,i}, \exp_{x_k}^{-1} \tilde{x}_{k,i} \right\rangle + \left\langle \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i}, \exp_{\tilde{x}_{k,i}}^{-1} x_k \right\rangle \geq d^2(x_k, \tilde{x}_{k,i}). \tag{3.7}$$

Combining (3.6) and (3.7), we obtain $d^2(x_k, \tilde{x}_{k,i}) \leq 0$. Therefore, $x_k = \tilde{x}_{k,i}$, which implies $v_k = 0$. □

The following lemma plays a crucial role in the analysis of Algorithm 1. It provides an estimate for the difference between a point transported back to the same tangent space along different paths on the manifold. See [14, Lemma 2] for its proof.

Lemma 3.1. *Let Ω be a compact set in $\mathscr{M}$. Then there exists a constant $C > 0$ such that, for any $x, y, z \in \Omega$, the following inequalities hold*

$$d(x,y) \leq \left\| \exp_x^{-1} z - \mathscr{P}_{y,x} \exp_y^{-1} z \right\| \leq C d(x,y).$$

Given a pair of bounded sequences $\{(u_k, v_{k,i})\}$, we use the notation $S(u_k, v_{k,i}) = o(d(u_k, v_{k,i}))$ to mean that if $\lim_{k\to\infty} d(u_k, v_{k,i(k)}) = 0$ (resp., if $\lim_{i\to\infty} d(u_k, v_{k,i}) = 0$), then

$$\lim_{k\to\infty} \frac{S(u_k, v_{k,i(k)})}{d(u_k, v_{k,i(k)})} = 0 \quad \left(\text{resp.}, \lim_{i\to\infty} \frac{S(u_k, v_{k,i})}{d(u_k, v_{k,i})} = 0 \right).$$

Clearly, for fixed k, the first order Taylor's expansion of f yields

$$f(v_{k,i}) = f(u_k) + \langle \operatorname{grad} f(u_k), \exp_{u_k}^{-1} v_{k,i} \rangle + S(u_k, v_{k,i}),$$

with $S(u_k, v_{k,i}) = o(d(u_k, v_{k,i}))$. Even though this relation is not immediate if k is not fixed and $i = i(k)$, we argue that it still holds. Indeed, let us consider $\gamma(t) = \exp_{u_k}(tv)$, with $v = \exp_{u_k}^{-1} v_{k,i} \in T_{u_k}\mathscr{M}$, where $\gamma(t)$ denotes the geodesic joining $u_k = \gamma(0)$ and $v_{k,i} = \gamma(1)$, and moreover $\psi : [0,1] \to \mathbb{R}$, $\psi(t) = f(\gamma(t))$. From chain rule, we obtain

$$\psi'(t) = \left\langle \operatorname{grad} f(\gamma(t)),\ \gamma'(t) \right\rangle_{\gamma(t)} = \left\langle\ \operatorname{grad} f(\gamma(t)),\ \mathrm{D}(\exp_{u_k})_{tv}[v] \right\rangle_{\gamma(t)}.$$

By applying the Mean Value Theorem, there exists $c_{k,i} := \gamma(\bar{t})$, $\bar{t} \in (0,1)$ such that

$$f(v_{k,i}) = f(u_k) + \langle \operatorname{grad} f(c_{k,i}), \mathrm{D}(\exp_{u_k})_{\bar{t}v}[v] \rangle. \tag{3.8}$$

Assuming that $\{(u_k, v_{k,i})\}$ is bounded and approaches zero, we clearly have that any accumulation point of u_k, consequently of $\{v_{k,i}\}$, is also an accumulation point of $\{c_{k,i}\}$. Moreover, it follows from (3.8) that

$$f(v_{k,i}) = f(u_k) + \langle \operatorname{grad} f(u_k), \exp_{u_k}^{-1} v_{k,i} \rangle + S(u_k, v_{k,i}),$$

where $S(u_k, v_{k,i}) = \langle \operatorname{grad} f(c_{k,i}), \mathrm{D}(\exp_{u_k})_{\bar{t}v}[v] \rangle - \langle \operatorname{grad} f(u_k), \exp_{u_k}^{-1} v_{k,i} \rangle$. By applying the Cauchy-Schwarz inequality to the last equality, we have

$$\|S(u_k, v_{k,i})\| \leq \|\operatorname{grad} f(c_{k,i})\| \|\mathrm{D}(\exp_{u_k})_{\bar{t}v}[v]\| + \|\operatorname{grad} f(u_k)\| \|\exp_{u_k}^{-1} v_{k,i}\|,$$

from which the claim follows.

The following lemma provides a basic inequality that will be used to simplify the proofs of our results.

Lemma 3.2. *Let a bounded sequence of elements $\{(x_k, \tilde{x}_{k,i}, \hat{x}_{k,i}, \alpha_{k,i})\}$ defined as in step 2 of Algorithm 1. Then, for every iteration pair (k,i), the following inequality holds*

$$F(\tilde{x}_{k,i}) \leq F(x_k) - \frac{1}{2\alpha_{k,i}} d^2(x_k, \tilde{x}_{k,i}) + R_{k,i},$$

where $R_{k,i} := S(x_k, \tilde{x}_{k,i}) = o(d(x_k, \tilde{x}_{k,i}))$. As a consequence, if (3.3) does not hold, then

$$R_{k,i} > \frac{1-\sigma}{2\alpha_{k,i}} d^2(x_k, \tilde{x}_{k,i}).$$

Proof. It follows from the Taylor's expansion for the function f and the definition of $R_{k,i}$ that

$$f(\tilde{x}_{k,i}) = f(x_k) + \langle \operatorname{grad} f(x_k), \exp_{x_k}^{-1} \tilde{x}_{k,i} \rangle + R_{k,i}. \tag{3.9}$$

Since $\tilde{x}_{k,i}$ satisfies (3.2) or, equivalently, is the solution of the proximal subproblem (3.1) with $x = x_k$ and $\alpha = \alpha_k$, we have

$$g(\tilde{x}_{k,i}) \leq g(x_k) + \frac{1}{2\alpha_{k,i}} d^2(x_k, \hat{x}_{k,i}) - \frac{1}{2\alpha_{k,i}} d^2(\tilde{x}_{k,i}, \hat{x}_{k,i}),$$

which combined with (3.9) and the fact that $F = f + g$ yields

$$\begin{aligned} F(\tilde{x}_{k,i}) &= f(\tilde{x}_{k,i}) + g(\tilde{x}_{k,i}) \\ &\leq F(x_k) + \langle \operatorname{grad} f(x_k), \exp_{x_k}^{-1} \tilde{x}_{k,i} \rangle + \frac{1}{2\alpha_{k,i}} \left(d^2(x_k, \hat{x}_{k,i}) - d^2(\tilde{x}_{k,i}, \hat{x}_{k,i}) \right) + R_{k,i}. \end{aligned} \tag{3.10}$$

Now, note that inequality (2.1) with $y_j = x_k$, $y_{j+1} = \hat{x}_{k,i}$ and $y_{j+2} = \tilde{x}_{k,i}$ implies that

$$\begin{aligned} d^2\left(x_k, \hat{x}_{k,i}\right) - d^2\left(\tilde{x}_{k,i}, \hat{x}_{k,i}\right) &\leq -d^2\left(\tilde{x}_{k,i}, x_k\right) + 2\left\langle \exp_{x_k}^{-1} \hat{x}_{k,i}, \exp_{x_k}^{-1} \tilde{x}_{k,i} \right\rangle \\ &= -d^2(\tilde{x}_{k,i}, x_k) - 2\alpha_{k,i} \left\langle \operatorname{grad} f\left(x_k\right), \exp_{x_k}^{-1} \tilde{x}_{k,i} \right\rangle, \end{aligned}$$

where the last relation is due to the fact that $\hat{x}_{k,i} = \exp_{x_k}\left(-\alpha_{k,i} \operatorname{grad} f(x_k)\right)$. Therefore, the first statement of the lemma follows by combining the last inequality with (3.10). The last statement of the lemma follows immediately from the first one and the assumption that (3.3) does not hold. □

Henceforth, we assume that the following assumption is satisfied.

(A1) The function $F = f + g$ is *1 − coercive* on $\mathscr{M}$.

Note that assumption **(A1)** implies that the level set M^c is compact and therefore guarantees that the sequence $\{x_k\}$ generated by the algorithm, has accumulation points.

The next lemma shows that if the current point is not critical, then the inner loop in Step 2 of Algorithm 1 is finite.

Lemma 3.3. *Assume that x_k generated by Algorithm 1 is not a critical point of Problem 1.1. Then, the inner loop in Algorithm 1 is finite.*

Proof. Assume by contradiction that the inner loop in Step 2 of Algorithm 1 does not terminate at some outer iteration k. Since $\tilde{x}_{k,i}$ solves the proximal subproblem (3.1), we have

$$g(\tilde{x}_{k,i}) + \frac{1}{2\alpha_{k,i}} d^2(\tilde{x}_{k,i}, \hat{x}_{k,i}) \leq g(\hat{x}_{k,i}) + \frac{1}{2\alpha_{k,i}} d^2(\hat{x}_{k,i}, \hat{x}_{k,i}) = g(\hat{x}_{k,i}). \tag{3.11}$$

For any $i \in \mathbb{N}$, let $u_{k,i} \in \partial g(\hat{x}_{k,i})$. Hence, $g(\hat{x}_{k,i}) \leq g(\tilde{x}_{k,i}) - \langle u_{k,i}, \exp_{\hat{x}_{k,i}}^{-1} \tilde{x}_{k,i} \rangle$, which, combined with (3.11) and the Cauchy-Schwarz inequality, implies that

$$\frac{1}{2\alpha_{k,i}} d^2(\tilde{x}_{k,i}, \hat{x}_{k,i}) \leq -\langle u_{k,i}, \exp_{\hat{x}_{k,i}}^{-1} \tilde{x}_{k,i} \rangle \leq \|u_{k,i}\| \| \exp_{\hat{x}_{k,i}}^{-1} \tilde{x}_{k,i} \|.$$

Since $d(\tilde{x}_{k,i}, \hat{x}_{k,i}) = \| \exp_{\hat{x}_{k,i}}^{-1} \tilde{x}_{k,i} \|$, the above inequality implies that

$$d(\tilde{x}_{k,i}, \hat{x}_{k,i}) \leq 2\alpha_{k,i} \|u_{k,i}\|. \tag{3.12}$$

Now, note that $\alpha_{k,i} = \tau^i \alpha_0^k \to 0$ as $i \to \infty$. Hence, $\hat{x}_{k,i} = \exp_{x_k}\left(-\alpha_{k,i}\operatorname{grad} f(x_k)\right)$ converges to x_k as $i \to \infty$. In particular, we have that $\{u_{k,i}\}_{i\in\mathbb{N}}$ is bounded, in view of Proposition 2.3. Thus, it follows from (3.12) that

$$\lim_{i\to\infty} \tilde{x}_{k,i} = \lim_{i\to\infty} \hat{x}_{k,i} = x_k. \tag{3.13}$$

Now, we claim that

$$\inf_{i\in\mathbb{N}} \frac{1}{\alpha_{k,i}} d(x_k, \tilde{x}_{k,i}) > 0. \tag{3.14}$$

Indeed, assume by contradiction that there exists a subsequence $\{i_l\}$ such that

$$\lim_{l\to\infty} \frac{1}{\alpha_{k,i_l}} d(x_k, \tilde{x}_{k,i_l}) = 0. \tag{3.15}$$

Since $\tilde{x}_{k,i_l}$ is the solution of the proximal inclusion (3.2) with $i = i_l$, we obtain

$$0 \in \frac{1}{2\alpha_{k,i_l}} \operatorname{grad} d^2(\tilde{x}_{k,i_l}, \hat{x}_{k,i_l}) + \partial g(\tilde{x}_{k,i_l}) = -\frac{1}{\alpha_{k,i_l}} \exp^{-1}_{\tilde{x}_{k,i_l}} \hat{x}_{k,i_l} + \partial g(\tilde{x}_{k,i_l}),$$

where the last equality is due to the last relation in Example 2.1. Thus we have

$$w_{k,i_l} := \frac{1}{\alpha_{k,i_l}} \exp^{-1}_{\tilde{x}_{k,i_l}} \hat{x}_{k,i_l} + \operatorname{grad} f(\tilde{x}_{k,i_l}) \in \operatorname{grad} f(\tilde{x}_{k,i_l}) + \partial g\left(\tilde{x}_{k,i_l}\right). \tag{3.16}$$

Let us now show that the sequence $\{w_{k,i_l}\}$ converges to zero as $l \to \infty$. Using the triangle inequality, we have

$$\begin{aligned} \|w_{k,i_l}\| &= \left\| \operatorname{grad} f(\tilde{x}_{k,i_l}) - \mathscr{P}_{x_k,\tilde{x}_{k,i_l}} \operatorname{grad} f(x_k) + \mathscr{P}_{x_k,\tilde{x}_{k,i_l}} \operatorname{grad} f(x_k) + \frac{1}{\alpha_{k,i_l}} \exp^{-1}_{\tilde{x}_{k,i_l}} \hat{x}_{k,i_l} \right\| \\ &\leq \left\| \operatorname{grad} f(\tilde{x}_{k,i_l}) - \mathscr{P}_{x_k,\tilde{x}_{k,i_l}} \operatorname{grad} f(x_k) \right\| + \frac{1}{\alpha_{k,i_l}} \left\| \mathscr{P}_{x_k,\tilde{x}_{k,i_l}} \left(\alpha_{k,i_l} \cdot \operatorname{grad} f(x_k)\right) + \exp^{-1}_{\tilde{x}_{k,i_l}} \hat{x}_{k,i_l} \right\|. \end{aligned} \tag{3.17}$$

Since, by (3.13), $\{\tilde{x}_{k,i}\}$ converges to x_k as $i \to \infty$, using the continuity of parallel transport and the continuity of $\operatorname{grad} f$, we conclude that

$$\zeta_{k,l} := \left\| \operatorname{grad} f(\tilde{x}_{k,i_l}) - \mathscr{P}_{x_k,\tilde{x}_{k,i_l}} \operatorname{grad} f(x_k) \right\| \overset{l\to\infty}{\longrightarrow} 0. \tag{3.18}$$

On the other hand, note that $\exp^{-1}_{x_k} \hat{x}_{k,i_l} = -\alpha_{k,i_l} \operatorname{grad} f(x_k)$, k is fixed, and $\{(\tilde{x}_{k,i}, \hat{x}_{k,i})\}_{i\in\mathbb{N}}$ is bounded, in view of (3.13). In particular, there exists a compact set Ω such that the triple $(\tilde{x}_{k,i}, x_k, \hat{x}_{k,i})$ belongs to Ω, for all $i \in \mathbb{N}$. Therefore, it follows from Lemma 3.1 with $x = \tilde{x}_{k,i_l}$, $y = x_k$ and $z = \hat{x}_{k,i_l}$, that there exists $C > 0$ such that

$$\begin{aligned} &\frac{1}{\alpha_{k,i_l}} \left\| \exp^{-1}_{\tilde{x}_{k,i_l}} \hat{x}_{k,i_l} + \mathscr{P}_{x_k,\tilde{x}_{k,i_l}} \left(\alpha_{k,i_l} \cdot \operatorname{grad} f(x_k)\right) \right\| \\ &= \frac{1}{\alpha_{k,i_l}} \left\| \exp^{-1}_{\tilde{x}_{k,i_l}} \hat{x}_{k,i_l} - \mathscr{P}_{x_k,\tilde{x}_{k,i_l}} \left(\exp^{-1}_{x_k} \hat{x}_{k,i_l}\right) \right\| \\ &\leq \frac{C}{\alpha_{k,i_l}} d(x_k, \tilde{x}_{k,i_l}). \end{aligned}$$

Combining the latter inequality with (3.17) and the definition of $\zeta_{k,l}$, we have

$$\left\| w_{k,i_l} \right\| \leq \zeta_{k,l} + \frac{C}{\alpha_{k,i_l}} d(x_k, \tilde{x}_{k,i_l}).$$

It follows from the above inequality, (3.15), and (3.18) that $\{w_{k,i_l}\}$ converges to zero as $l \to \infty$. Hence, since, in view of (3.13), $\lim_{i\to\infty} \tilde{x}_{k,i} = x_k$, we have from (3.16), the continuity of grad f, and the closed graph property of ∂g (see Remark 2.1), that $0 \in \text{grad } f(x_k) + \partial g(x_k)$, which implies that x_k is a critical point of Problem 1.1, contradicting the assumption of the lemma. Therefore, the claim is proved, i.e., (3.14) holds. On the other hand, since we are assuming that the inner loop of Algorithm 1 does not stop, we have from Lemma 3.2 that there exists $R_{k,i} = o(d(x_k, \tilde{x}_{k,i}))$ such that

$$\frac{R_{k,i}}{d(x_k, \tilde{x}_{k,i})} > \frac{1-\sigma}{2\alpha_{k,i}} d(x_k, \tilde{x}_{k,i}).$$

In view of (3.13), $\lim_{i\to\infty} \tilde{x}_{k,i} = x_k$, and $\sigma \in (0,1)$, we obtain $\lim_{i\to\infty} \frac{1}{\alpha_{k,i}} d(x_k, \tilde{x}_{k,i}) = 0$, contradicting (3.14). Therefore, we conclude that the inner loop must indeed be finite, and the lemma is proved. □

4. Convergence Analysis

In this section, we carry out the convergence analysis of Algorithm 1 and prove the main convergence results of this paper, namely, Theorems 4.1 and 4.2, which show, respectively, that every accumulation point of the sequence $\{x_k\}$ is a critical point to problem 1.1 and that the whole sequence converges under a special property. To this end, we assume that Algorithm 1 generates an infinite sequence $\{x_k\}$. To establish the aforementioned results, we first state and prove three technical lemmas. First, we analyze the sequence of the distance between consecutive iterates $\{d(x_k, x_{k+1})\}$ and show, in particular, that it converges to zero. Second, we show that, along a convergent subsequence of $\{x_k\}$, the sequence $\{d(x_k, z_k)/\alpha_k\}$ converges to zero even if $\{\alpha_k\}$ converges to zero, where $z_k := \tilde{x}_{k,i_k-1}$ is the second-to-last iterate of the inner procedure used to update x_k. Finally, we consider a special sequence of "subgradients" of F and show an important relation that will be used to establish our main results.

Lemma 4.1. *The following inequalities hold*

(i) $\sum_{k=0}^{+\infty} \frac{1}{\alpha_k} d^2(x_k, x_{k+1}) < \infty$.
(ii) $\lim_{k\to\infty} d(x_k, x_{k+1}) = 0$.

Proof. (i) Using the acceptance criterion (3.3) and the definitions of α_k and x_{k+1}, we have

$$F(x_{k+1}) \leq F(x_k) - \frac{\sigma}{2\alpha_k} d^2(x_k, x_{k+1}), \qquad \forall\, k \in \mathbb{N}. \tag{4.1}$$

Hence, by Remark 3.1, the sequence of functional values is monotonically decreasing. Moreover, since **(A1)** holds, then the level set $M^{F(x_0)}$ is compact. Therefore $\{x_k\}$ has accumulation points. In particular, we have $-\infty < F^* := \lim_{j\to\infty} F(x_j) \leq F(x_k), \forall\, k \in \mathbb{N}$. Then, it follows from (4.1) that

$$\sum_{k=0}^{K} \frac{\sigma}{2\alpha_k} d^2(x_k, x_{k+1}) \leq \sum_{k=0}^{K} (F(x_k) - F(x_{k+1})) = F(x_0) - F(x_K),$$

which implies that

$$\sum_{k=0}^{+\infty} \frac{\sigma}{2\alpha_k} d^2(x_k, x_{k+1}) \leq F(x_0) - F^* < \infty,$$

proving statement (i).

(ii) This statement follows immediately from (i) and the fact that $\frac{1}{\alpha_k} \geq \frac{1}{\alpha_{\max}}$ for all $k \in \mathbb{N}$. □

Lemma 4.2. *Let $\{x_k\}_{k\in\mathcal{K}}$ be a convergent subsequence of $\{x_k\}$ and assume that $\{\alpha_k\}_{k\in\mathcal{K}}$ converges to zero. Then,*

$$\lim_{k\to\infty,k\in\mathcal{K}} d(x_k,z_k) = 0, \qquad \lim_{k\to\infty,k\in\mathcal{K}} \frac{1}{\bar{\alpha}_k} d(x_k,z_k) = 0,$$

where

$$\bar{\alpha}_k := \alpha_{k,i_k-1} = \frac{\alpha_k}{\tau}, \qquad z_k := \tilde{x}_{k,i_k-1}, \qquad \forall k \in \mathcal{K}.$$

Proof. First note that $\{\bar{\alpha}_k\}_{k\in\mathcal{K}}$ converges to zero. Using the definition of $\bar{\alpha}_k$ and the fact that z_k is the solution of the proximal subproblem at step 2 of Algorithm 1 with $i = i_k - 1$, we have

$$g(z_k) + \frac{1}{2\bar{\alpha}_k} d^2(z_k,\hat{z}_k) \leq g(x_k) + \frac{1}{2\bar{\alpha}_k} d^2(x_k,\hat{z}_k), \qquad \forall k \in \mathcal{K},$$

where $\hat{z}_k := \hat{x}_{k,i_k-1}$. This inequality can be rewritten as

$$g(z_k) - g(x_k) \leq \frac{1}{2\bar{\alpha}_k}[d^2(x_k,\hat{z}_k) - d^2(z_k,\hat{z}_k)]. \tag{4.2}$$

On the other hand, it follows from Proposition 2.1 with $y_j = x_k$, $y_{j+1} = \hat{z}_k$ and $y_{j+2} = z_k$ that

$$\begin{aligned} d^2(x_k,\hat{z}_k) - d^2(z_k,\hat{z}_k) &\leq -d^2(z_k,x_k) + 2\left\langle \exp_{x_k}^{-1}\hat{z}_k, \exp_{x_k}^{-1} z_k\right\rangle \\ &= -d^2(x_k,z_k) - 2\bar{\alpha}_k\left\langle \operatorname{grad} f(x_k), \exp_{x_k}^{-1} z_k\right\rangle, \end{aligned} \tag{4.3}$$

where the last relation is due to the fact that $\hat{z}_k = \exp_{x_k}(-\bar{\alpha}_k \operatorname{grad} f(x_k))$. It follows by (4.2), (4.3), the Cauchy-Schwartz inequality, and by considering $u_k \in \partial g(x_k)$ for any $k \in \mathcal{K}$, that

$$\begin{aligned} \frac{1}{2\bar{\alpha}_k} d^2(x_k,z_k) &\leq g(x_k) - g(z_k) - \left\langle \operatorname{grad} f(x_k), \exp_{x_k}^{-1} z_k\right\rangle \\ &\leq -\langle u_k, \exp_{x_k}^{-1} z_k\rangle - \left\langle \operatorname{grad} f(x_k), \exp_{x_k}^{-1} z_k\right\rangle \\ &= -\left\langle u_k + \operatorname{grad} f(x_k), \exp_{x_k}^{-1} z_k\right\rangle \\ &\leq \|u_k + \operatorname{grad} f(x_k)\| \|\exp_{x_k}^{-1} z_k\|, \end{aligned} \tag{4.4}$$

where the second inequality is due to $g(z_k) \geq g(x_k) + \langle u_k, \exp_{x_k}^{-1} z_k\rangle$ for any $k \in \mathcal{K}$. Since $d(x_k,z_k) = \|\exp_{x_k}^{-1} z_k\|$, it follows from (4.4) that

$$d(x_k,z_k) \leq 2\bar{\alpha}_k \|u_k + \operatorname{grad} f(x_k)\|, \qquad \forall k \in \mathcal{K}.$$

Now, since $\{x_k\}_{k\in\mathcal{K}}$ is convergent, we have that $\{\operatorname{grad} f(x_k)\}_{k\in\mathcal{K}}$ and $\{u_k\}_{k\in\mathcal{K}}$ are bounded, in view of the continuity of grad f, the fact that $u_k \in \partial g(x_k)$, and Proposition 2.3. Thus, since $\{\bar{\alpha}_k\}_{k\in\mathcal{K}}$ converges to zero, it follows from the above inequality that $\{d(x_k,z_k)\}_{k\in\mathcal{K}}$ converges to zero. Now, since z_k does not satisfy the acceptance criterion (3.3) and $\{(x_k,z_k)\}_{k\in\mathcal{K}}$ is bounded, it follows from Lemma 3.2 with $i = i_k - 1$ and the definition of z_k that there exists $\bar{R}_k := R_{k,i_k-1} = o(d(x_k,z_k))$ such that

$$\frac{\bar{R}_k}{d(x_k,z_k)} > \frac{1-\sigma}{2\bar{\alpha}_k} d(x_k,z_k).$$

Therefore, since $\{d(x_k,z_k)\}_{k\in\mathcal{K}}$ converges to zero, the left hand side of the above inequality also goes to zero, and then the proof of the lemma follows from the above inequality by noting that $\sigma \in (0,1)$. □

Lemma 4.3. *Let sequences* $\{(x_k,\tilde{x}_{k,i},\hat{x}_{k,i})\}$ *and* $\{\alpha_{k,i}\}$ *be generated by Step 2 of Algorithm 1 and define*

$$s_{k,i} := \operatorname{grad} f(\tilde{x}_{k,i}) + \frac{1}{\alpha_{k,i}} \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i}. \tag{4.5}$$

Then, for every pair (k,i), *there hold* $s_{k,i} \in \operatorname{grad} f(\tilde{x}_{k,i}) + \partial g\left(\tilde{x}_{k,i}\right)$ *and*

$$\|s_{k,i}\| \leq \left\|\operatorname{grad} f(\tilde{x}_{k,i}) - \mathscr{P}_{x_k,\tilde{x}_{k,i}} \operatorname{grad} f(x_k)\right\| + \frac{C}{\alpha_{k,i}} d(x_k,\tilde{x}_{k,i}), \tag{4.6}$$

for some scalar $C > 0$ *that does not depend on* (k,i).

Proof. Recall that $\hat{x}_{k,i} = \exp_{x_k}\left(-\alpha_{k,i} \operatorname{grad} f(x_k)\right)$. By the optimality condition (3.2) and the last relation of Example 2.1, we have

$$0 \in \frac{1}{2\alpha_{k,i}} \operatorname{grad} d^2\left(\tilde{x}_{k,i},\hat{x}_{k,i}\right) + \partial g\left(\tilde{x}_{k,i}\right) = -\frac{1}{\alpha_{k,i}} \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i} + \partial g\left(\tilde{x}_{k,i}\right).$$

Thus, we have

$$s_{k,i} = \operatorname{grad} f(\tilde{x}_{k,i}) + \frac{1}{\alpha_{k,i}} \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i} \in \operatorname{grad} f(\tilde{x}_{k,i}) + \partial g\left(\tilde{x}_{k,i}\right),$$

proving the first statement of the lemma. Now, using the definition of $s_{k,i}$ and the triangle inequality, we have

$$\begin{aligned}
\|s_{k,i}\| &= \left\|\operatorname{grad} f(\tilde{x}_{k,i}) + \frac{1}{\alpha_{k,i}} \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i}\right\| \\
&= \left\|\operatorname{grad} f(\tilde{x}_{k,i}) - \mathscr{P}_{x_k,\tilde{x}_{k,i}} \operatorname{grad} f(x_k) + \mathscr{P}_{x_k,\tilde{x}_{k,i}} \operatorname{grad} f(x_k) + \frac{1}{\alpha_{k,i}} \exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i}\right\| \\
&\leq \left\|\operatorname{grad} f(\tilde{x}_{k,i}) - \mathscr{P}_{x_k,\tilde{x}_{k,i}} \operatorname{grad} f(x_k)\right\| + \frac{1}{\alpha_{k,i}} \left\|\exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i} + \mathscr{P}_{x_k,\tilde{x}_{k,i}}\left(\alpha_{k,i} \operatorname{grad} f(x_k)\right)\right\|.
\end{aligned} \tag{4.7}$$

Now, note that from Remark 3.1, we have $F(x_k) \leq F(x_0)$ for all k, and from Assumption **(A1)**, we have that $\{x_k\}$ is bounded. From the continuity of the exponential map and the gradient of f, the definition $\hat{x}_{k,i} = \exp_{x_k}\left(-\alpha_{k,i} \operatorname{grad} f(x_k)\right)$, and the fact that both sequences $\{\alpha_{k,i}\}$ and $\{x_k\}$ are bounded, we conclude that $\{\hat{x}_{k,i}\}$ is also bounded. On the other hand, similarly to the first part of the proof of Lemma 3.3, one may show that (3.12) holds, i.e., $d(\tilde{x}_{k,i},\hat{x}_{k,i}) \leq 2\,\alpha_{k,i}\,\|u_{k,i}\|$, where $u_{k,i} \in \partial g(\hat{x}_{k,i})$, for every pair (k,i). Since $\{\hat{x}_{k,i}\}$ is bounded, Proposition 2.3 implies that $\{u_{k,i}\}$ is also bounded. Hence, it follows from the latter inequality and the boundedness of $\{\alpha_{k,i}\}$ and $\{\hat{x}_{k,i}\}$ that $\{\tilde{x}_{k,i}\}$ is also bounded. We then conclude that there exists a compact set Ω such that the triple $(\tilde{x}_{k,i},x_k,\hat{x}_{k,i})$ belongs to Ω, for all pairs of indices (k,i). Hence, it follows from Lemma 3.1 with $x = \tilde{x}_{k,i}$, $y = x_k$, and $z = \hat{x}_{k,i}$ that there exists $C > 0$ such that

$$\begin{aligned}
\frac{1}{\alpha_{k,i}} \left\|\exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i} + \mathscr{P}_{x_k,\tilde{x}_{k,i}}\left(\alpha_{k,i} \operatorname{grad} f(x_k)\right)\right\| &= \frac{1}{\alpha_{k,i}} \left\|\exp_{\tilde{x}_{k,i}}^{-1} \hat{x}_{k,i} - \mathscr{P}_{x_k,\tilde{x}_{k,i}}\left(\exp_{x_k}^{-1} \hat{x}_{k,i}\right)\right\| \\
&\leq \frac{C}{\alpha_{k,i}} d(x_k,\tilde{x}_{k,i}).
\end{aligned}$$

Therefore, the second statement of the lemma follows by combining the latter inequality and (4.7). □

In the following, we present the main convergence results of this paper. The first one shows that any accumulation points of the sequence $\{x_k\}$ is critical to Problem 1.1. The second result establishes the convergence of the whole sequence $\{x_k\}$ under the assumption that the objective function has an isolated critical point which is a accumulation point of $\{x_k\}$.

Theorem 4.1. *The sequence $\{x_k\}$ generated by Algorithm 1 is bounded, and every accumulation point of it is a critical point of Problem 1.1.*

Proof. First note from Remark 3.1 that $\{F(x_k)\}$ is decreasing and hence $\{x_k\}$ is contained in $M^{F(x_0)}$, the first sub-level set of F, which is a compact set in view of **(A1)**. Thus, $\{x_k\}$ is bounded. Now, let a subsequence $\{x_k\}_{k\in\mathcal{K}}$ converge to an accumulation point x^*, and let $\{\alpha_k\}$ be the sequence of step sizes generated by Algorithm 1. We continue the proof by analyzing two cases, depending on whether the subsequence $\{\alpha_k\}_{k\in\mathcal{K}}$ converges to zero or not. First, assume that $\{\alpha_k\}_{k\in\mathcal{K}}$ converges to zero and define

$$z_k = \tilde{x}_{k,i_k-1}, \quad \bar{\alpha}_k = \alpha_{k,i_k-1} = \frac{\alpha_k}{\tau}, \quad \hat{z}_k = \hat{x}_{k,i_k-1} = \exp_{x_k}(-\bar{\alpha}_k \operatorname{grad} f(x_k)).$$

Considering $s_{k,i}$ as in Lemma 4.3, it follows from (4.5) and (4.6) that

$$s_{k,i_k-1} = \operatorname{grad} f(z_k) + \exp_{z_k}^{-1} \hat{z}_k \in \operatorname{grad} f(z_k) + \partial g(z_k), \tag{4.8}$$

$$\|s_{k,i_k-1}\| \leq \|\operatorname{grad} f(z_k) - \mathscr{P}_{x_k,z_k} \operatorname{grad} f(x_k)\| + \frac{C}{\bar{\alpha}_k} d(x_k,z_k). \tag{4.9}$$

Now, since $\{x_k\}_{k\in\mathcal{K}} \to x^*$ and $\{\bar{\alpha}_k\}_{k\in\mathcal{K}} \to 0$, Lemma 4.2 together with the definition of $\hat{z}_k$ implies that $\{d(x_k,z_k)\}_{k\in\mathcal{K}}$ and $\{\frac{1}{\bar{\alpha}_k} d(x_k,z_k)\}_{k\in\mathcal{K}}$ both converge to zero, and that $\{z_k\}_{k\in\mathcal{K}}$ and $\{\hat{z}_k\}_{k\in\mathcal{K}}$ both converge to x^*. Hence, it follows from (4.8) and (4.9) that $\{s_{k,i_k-1}\}_{k\in\mathcal{K}}$ converges to zero and that $0 \in \operatorname{grad} f(x^*) + \partial g(x^*)$, in view of the continuity of $\operatorname{grad} f$ and the closed graph property of ∂g (see Remark 2.1). Consequently, the statement of the theorem follows if the $\{\alpha_k\}_{k\in\mathcal{K}}$ converges to zero.

Now, we assume that $\{\alpha_k\}_{k\in\mathcal{K}}$ does not converge to zero. In this case, using Lemma 4.3 with $i = i_k$, and recalling that $x_{k+1} = \tilde{x}_{k,i_k}$, $\alpha_k = \alpha_{k,i_k}$ and $\hat{x}_{k,i_k} = \exp_{x_k}(-\alpha_k \operatorname{grad} f(x_k))$, we have

$$s_{k,i_k} = \operatorname{grad} f(x_{k+1}) - \frac{1}{2\alpha_k} \operatorname{grad} d^2(x_{k+1}, \hat{x}_{k,i_k}) \in \operatorname{grad} f(x_{k+1}) + \partial g(x_{k+1}), \tag{4.10}$$

$$\|s_{k,i_k}\| \leq \left\|\operatorname{grad} f(x_{k+1}) - \mathscr{P}_{x_k,x_{k+1}} \operatorname{grad} f(x_k)\right\| + \frac{C}{\alpha_k} d(x_k,x_{k+1}). \tag{4.11}$$

Since, by Lemma 4.1(ii), the sequence $\{d(x_k,x_{k+1})\}_{k\in\mathcal{K}}$ converges to zero, using (4.10), (4.11), together with the facts that $\{x_k\}_{k\in\mathcal{K}} \to x^*$ and $\{\alpha_k\}_{k\in\mathcal{K}} \not\to 0$, we deduce that a subsequence of $\{s_{k,i_k}\}_{k\in\mathcal{K}}$ converges to zero and that $0 \in \operatorname{grad} f(x^*) + \partial g(x^*)$. Therefore, x^* is a critical point of Problem 1.1, and the proof of the theorem is complete. □

In view of Lemma 4.1, we see that the next result may be proved in a manner similar to [28]; see also [23]. However, we present a detailed proof for the sake of completeness.

Theorem 4.2. *Let $\bar{x}$ be an accumulation point of the sequence $\{x_k\}$, which is an isolated critical point of the objective function F. Then the entire sequence $\{x_k\}$ converges to $\bar{x}$.*

Proof. It follows from Theorem 4.1 that $\{x_k\}$ is bounded and every accumulation point belongs to $\operatorname{crit}(F)$, the set of critical points of F. Now, since $\bar{x}$ is an isolated critical point for F, there exists a sufficiently small $\delta > 0$ such that $\bar{x}$ is the only critical point of F in $\mathscr{B}(\bar{x}, 4\delta)$. Consider the sets $S := \operatorname{crit}(F) \setminus \{\bar{x}\}$ and $S_\delta := \{x \in \mathscr{M} : d(x,S) \leq \delta\}$. We claim that there exists $k_0 \in \mathbb{N}$ such that $x_k \in \mathscr{B}(\bar{x},\delta) \cup S_\delta$, for all $k \geq k_0$. Indeed, if there is no such a $k_0 \in \mathbb{N}$, then we may construct a subsequence $\{x_k\}_{k\in\mathscr{K}} \subset (\mathscr{B}(\bar{x},\delta) \cup S_\delta)^C$. Since $\{x_k\}_{k\in\mathscr{K}}$ is bounded and every accumulation point belongs to $\operatorname{crit}(F)$, this implies that there exists a critical point $\hat{x}$ of F such that $\min\{d(\hat{x},\bar{x}), d(\hat{x},S)\} \geq \delta$ which contradicts the fact that $\operatorname{crit}(F) = S \cup \{\bar{x}\}$. Hence, the claim is true. Now, note that Lemma 4.1(ii) implies that $d(x_k, x_{k+1}) \to 0$, and then there exists $k_1 \geq k_0$ such that $d(x_k, x_{k+1}) \leq \delta$ for every $k \geq k_1$. Since $\bar{x}$ is an accumulation point of the sequence $\{x_k\}$, there exists a subsequence $\{x_k\}_{k\in\mathscr{K}} \subset \mathscr{B}(\bar{x},\delta)$. We may assume without loss of generality that $k \geq k_1$ for any $k \in \mathscr{K}$. We argue that for any $k \in \mathbb{N}$ such that $k \geq k_1$ and $x_k \in \mathscr{B}(\bar{x},\delta)$, we have that x_{k+1} also belongs to $\mathscr{B}(\bar{x},\delta)$. Indeed, let an arbitrary $x^* \in S$. Since $\bar{x}$ is the only critical point of F in $\mathscr{B}(\bar{x}, 4\delta)$, and $\max\{d(x_k,\bar{x}), d(x_k,x_{k+1})\} \leq \delta$, by applying the triangle inequality twice, we obtain

$$\begin{aligned} d(x_{k+1},x^*) &\geq d(x^*,\bar{x}) - d(\bar{x},x_k) - d(x_k,x_{k+1}) \\ &\geq 4\delta - \delta - \delta = 2\delta. \end{aligned}$$

Hence, since x^* is an arbitrary element in S, we conclude $d(x_{k+1},S) \geq 2\delta$. On the other hand, $x_{k+1} \in \mathscr{B}(\bar{x},\delta) \cup S_\delta$, because $k \geq k_1 \geq k_0$. Thus, we must have that $x_{k+1} \in \mathscr{B}(\bar{x},\delta)$. We then conclude that for any $k \geq k_1$, we have $x_k \in \mathscr{B}(\bar{x},\delta)$. Therefore, since every accumulation point of $\{x_k\}$ is critical for Problem (1.1) and $\bar{x}$ is the only critical point in $\mathscr{B}(\bar{x},\delta)$, we conclude that the entire sequence $\{x_k\}$ must converge to $\bar{x}$, proving the theorem. □

Remark 4.1. Note that the last theorem encompasses the particular case where the problem admits a finite number of critical points (see [24, Theorem 14.1.5]). It is worth emphasizing that the authors in [17] explored this result within the proximal gradient model in the Euclidean setting.

5. Conclusion

In this work, we extended a monotone proximal gradient method from the Euclidean setting to the context of Hadamard manifolds, focusing on solving nonsmooth and nonconvex optimization problems of the form $F = f + g$, where f is smooth, possibly nonconvex, and g is convex, possibly nonsmooth. The proximal gradient mapping is defined by solving a subproblem on the manifold, leveraging its geometric structure. We established the convergence of the proposed methods to critical points of the objective function F. The convergence analysis takes advantage of the special geometry of Hadamard manifolds, allowing applicability in settings where the Lipschitz continuity of the gradient of f is not required. We presented global convergence results under the assumption of the existence of an isolated critical point of the objective function.

Acknowledgments
The first author was supported in part by CNPq grants 314106/2020-0. The second author was supported in part by Capes/DS 5443857.

REFERENCES

[1] P.-A. Absil, R. Mahony, R. Sepulchre, Optimization algorithms on matrix manifolds, Princeton University Press, Princeton, 2008.

[2] Q.H. Ansari, F. Babu, Proximal point algorithm for inclusion problems in Hadamard manifolds with applications, Optim. Lett. 15 (2019), 901–921.

[3] A. Beck, M. Teboulle, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, IEEE Trans. Signal Process. 18 (11) (2009), 2419–2434.

[4] A. Beck, M. Teboulle, Gradient-based algorithms with applications to signal-recovery problems, Convex Optimization in Signal Processing and Communication, pp. 42-88, Cambridge University Press, Cambridge, 2010.

[5] G.C. Bento, O. Ferreira, J.G. Melo, Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds, J. Optim. Theory Appl. 173 (2017), 548–562.

[6] G.C. Bento, O. Ferreira, P. Oliveira, Local convergence of the proximal point method for a special class of nonconvex functions on Hadamard manifolds, Nonlinear Anal. 73 (2010), 564–572.

[7] G.C. Bento, O. Ferreira, P. Oliveira, Proximal point method for a special class of nonconvex functions on Hadamard manifolds, Optimization 64 (2015), 289–319.

[8] G.C. Bento, J.G. Melo, Subgradient method for convex feasibility on Riemannian manifolds, J. Optim. Theory Appl. 152 (2012), 773–785.

[9] R. Bergmann, H. Jasa, P. John, M. Pfeffer, The intrinsic Riemannian proximal gradient method for nonconvex optimization, arXiv:2506.09775v1 (2025).

[10] N. Boumal, An introduction to optimization on smooth manifolds, Cambridge Academic Press, Cambridge, 2023.

[11] S. Chen, S. Ma, A. Man-Cho So, T. Zhang, Proximal gradient method for nonsmooth optimization over the Stiefel manifold, SIAM J. Optim. 30 (2020), 210–239.

[12] J.X. Cruz Neto, L.L. de Lima, P.R. Oliveira, Geodesic algorithms in Riemannian geometry, Balkan J. Geom. Appl. 3 (1998), 89–100.

[13] M.P. do Carmo, Riemannian Geometry, Birkhäuser, Boston, 1992.

[14] S. Feng, W. Huang, L. Song, S. Ying, T. Zeng, Proximal gradient method for nonconvex and nonsmooth optimization on Hadamard manifolds, Optim. Lett. 16 (2022), 2277–2297.

[15] O.P. Ferreira, P.R. Oliveira, Proximal point algorithm on Riemannian manifolds, Optimization 51 (2002), 257–270.

[16] O.P. Ferreira, P.R. Oliveira, Subgradient algorithm on Riemannian manifolds, J. Optim. Theory Appl. 97 (1998), 93–104.

[17] M. Fukushima, H. Mine, A generalized proximal point algorithm for certain nonconvex minimization problems, Int. J. Syst. Sci. 12 (1981), 989–1000.

[18] S. Hosseini, M.S. Pouryayevali, Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds, Nonlinear Anal. 74 (2011), 3884–3895.

[19] W. Huang, K. Wei, Riemannian proximal gradient methods, Math. Program. 194 (2022), 371–413.

[20] C. Kanzow, P. Mehlitz, Convergence properties of monotone and nonmonotone proximal gradient methods revisited, J. Optim. Theory Appl. 195 (2022), 624–646.

[21] H. Li, Z. Lin, C. Fang, Accelerated Optimization for Machine Learning: First-order Algorithms, Springer, Singapore, 2020.

[22] X. Li, The proximal gradient method for composite optimization problems on Riemannian manifolds, Mathematics 12 (2024), 1–15.

[23] R.R. Meyer, Sufficient conditions for the convergence of monotonic mathematical programming algorithms, J. Comput. Syst. Sci. 12 (1976), 108–121.

[24] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear equations in Several Variables, Academic Press, USA, 1971.

[25] T. Sakai, Riemannian geometry, Translations of Mathematical Monographs 149, American Mathematical Society, USA, 1996.

[26] S. Sra, H. Zhang, First-order methods for geodesically convex optimization, JMLR Workshop and Conference Proceedings 49 (2016), 1–22.
[27] C. Udriste, Convex functions and optimization algorithms on Riemannian manifolds, Mathematics and its Applications 297, Kluwer Academic, NL, 1994.
[28] Y. Yang, Globally convergent optimization algorithms on Riemannian manifolds: Uniform framework for unconstrained and constrained optimization, J. Optim. Theory Appl. 132 (2007), 245–265.