This paper will appear in a forthcoming issue of the SIAM Journal on Control and Optimization

ON THE CONVERGENCE OF GRADIENT DESCENT FOR FINDING THE RIEMANNIAN CENTER OF MASS*

BIJAN AFSARI †, ROBERTO TRON ‡, AND RENÉ VIDAL †

Abstract. We study the problem of finding the global Riemannian center of mass of a set of data points on a Riemannian manifold. Specifically, we investigate the convergence of constant step-size gradient descent algorithms for solving this problem. The challenge is that often the underlying cost function is neither globally differentiable nor convex, and despite this one would like to have guaranteed convergence to the global minimizer. After some necessary preparations we state a conjecture which, we argue is the best convergence condition (in a specific described sense) that one can hope for. The conjecture specifies conditions on the spread of the data points, step-size range, and the location of the initial condition (i.e., the region of convergence) of the algorithm. These conditions depend on the topology and the curvature of the manifold and can be conveniently described in terms of the injectivity radius and the sectional curvatures of the manifold. For 2-dimensional manifolds of nonnegative curvature and manifolds of constant nonnegative curvature (e.g., the sphere in $\mathbb{R}^n$ and the rotation group in $\mathbb{R}^3$) we show that the conjecture holds true. For more general manifolds we prove convergence results which are weaker than the conjectured one (but still superior over the available results). We also briefly study the effect of the configuration of the data points on the speed of convergence. Finally, we study the global behavior of the algorithm on certain manifolds proving (generic) convergence of the algorithm to a local center of mass with an arbitrary initial condition. An important aspect of our presentation is our emphasize on the effect of curvature and topology of the manifold on the behavior of the algorithm.

Key words. Riemannian center of mass, Fréchet mean, Riemannian mean, Riemannian average, gradient descent, spherical geometry, spherical trigonometry, comparison theorems, gradient descent, convergence analysis, global convergence

AMS subject classifications. Primary 53C20, 90C25, 90C26; Secondary 62H11, 68U05, 92C55

1. Introduction.

The (global) Riemannian center of mass (a.k.a. Riemannian mean or average or Fréchet mean)[1] of a set of data points $\{x_i\}_{i=1}^N$ in a Riemannian manifold M is defined as the set of points which minimize the sum of squares of geodesic distances to the data points. This notion and its variants have a long history with several applications in pure mathematics (see e.g., [8, 25, 28, 30] and also [3] for a brief history and some new results). More recently, statistical analysis on Riemannian manifolds and, in particular, the Riemannian center of mass have found applications in many applied fields. These include fields such as computer vision (see e.g., [48, 47]), statistical analysis of shapes (see e.g., [29, 33, 10, 24, 9]), medical imaging (see e.g., [21, 38]), and sensor networks (see e.g., [46, 43]) and many other general data analysis applications (see e.g., [31, 2, 12, 37]). In these applied settings one often needs to numerically locate or compute the Riemannian center of mass of a set of data points lying on a Riemannian manifold.

*This work supported in part by grants NSF CAREER # 0447739, NSF CNS # 0834470, ONR # N00014-05-10836, ONR # N00014-09-1-0084, and # NSF ECCS-0941463.

†Center for Imaging Science, The Johns Hopkins University, Baltimore, MD, USA (email: {bijan, rvidal}@ cis.jhu.ed).

‡GRASP Lab, University of Pennsylvania, Philadelphia, PA, USA (email: tron@seas.upenn.edu). This work was done while Roberto Tron was with the Center for Imaging Science at Johns Hopkins.

[1] For most of this paper we are mainly interested in the "global" Riemannian center of mass, hence in reference to it very often we drop the term "global" and simply use "Riemannian center of mass," etc. If need arises we explicitly use the term "local" in reference to a center which is not global, necessarily (see Definition 2.5).

If the data points are localized enough, then their global Riemannian center of mass is a unique point $\bar{x} \in M$, which is also close to the data points. One can think of locating $\bar{x}$ using a *constant step-size* (intrinsic) gradient descent algorithm, which is the most popular and easiest version of gradient descent method. The main challenge, here, is that the underlying cost function is usually neither globally differentiable [2] nor globally convex on the manifold.[3] In fact, as it can be shown by simple examples, the cost function can have local minimizers, which are not of interest and should be avoided. Nevertheless, we expect and hope that if the algorithm is initialized *close enough* to the (unknown) global Riemannian center of mass $\bar{x}$ and the step-size is *small enough*, then the algorithm should converge to the center. One would like to have the step-size small enough so that the cost is reduced at each step and at the same time the iterates do not leave a neighborhood around $\bar{x}$ in which $\bar{x}$ is the only zero of the gradient of the cost function (recall that a gradient descent algorithm, at best, can only locate a zero of the gradient vector field of the cost function). On the other hand, one would like to have large enough step-size so that the convergence is fast. The interplay between these three constraints is important in determining the conditions guaranteeing convergence as well as the speed of convergence. The goal of this paper is to give accurate conditions that guarantee convergence of the constant step-size gradient algorithm to the global Riemannian center of mass of a set of data points.

1.1. Outline. In §2, we first briefly give the necessary backgrounds on the Riemannian center of mass and the gradient descent algorithm for finding it, these include the notions of convex functions and sets in §2.1.2, differentiability and convexity properties of the Riemannian distance function and bounds on its Hessian in §2.1.3, Riemannian center of mass and its properties in §2.1.4, a general convergence theorem for gradient descent in §2.1.5, a convergence theorem estimating the speed of convergence and the best step-size in §2.1.6, a comment on the role of contraction mapping in the convergence of the algorithm in §2.1.7. Following that, in §2.2, we state Conjecture 2.15 in which we specify the *best* "condition for convergence" one can hope for (in a sense to be described). Specifically, we specify a bound on the radius of any ball around the data points in which the algorithm can be initialized together with an interval of allowable step-sizes so that the algorithm converges to the global center of mass $\bar{x}$. The significant point is that for convergence, the radius of the ball does not need to be any smaller than what ensures existence and uniqueness of the center. Moreover, the step-size can be chosen equal to the best (in a specific described sense) step-size under the *extra* assumption that the iterates stay in that ball; [4] and it is conjectured that, indeed, the iterates stay in the ball. [5] Knowing the conjecture helps us to compare and evaluate the existing results mentioned in §2.3 as well as the results derived in this paper. In §3 (Theorem 3.7), we prove Conjecture 2.15 for the case of manifolds of *constant nonnegative* curvature as well as 2-dimensional mani-

[2] For us global differentiability means differentiability everywhere on the manifold; however, we use the term "global" to remind ourselves that our functions of interest (e.g., the Riemannian distance from a point) may lose differentiability at faraway distances.

[3] In fact, it is well known that on a compact Riemannian manifold the only globally continuous convex functions are constant functions (see Theorem 2.2 and [51]).

[4] This step-size, in general, depends on an upper bound on the sectional curvatures of the manifold and the radius of the mentioned ball. However, interestingly, for a manifold of nonnegative curvature it is simply 1 (see Conjecture 2.15).

[5] The main challenge in proving this conjecture is to prove that the iterates stay in the ball containing the data points.

folds of nonnegative curvature. In our developments in this section, we first prove comparison Theorem 3.1 in §3.1. This comparison result (which differs from standard comparison theorems in some aspects) most likely has been known among geometers, but we could find neither its statement nor a proof for it in the literature. In §3.2 we make sense of an intuitive notion of Riemannian convex combination of a set of data points in the mentioned manifolds and we explore its relation to the convex hull of the data points. These prepare us to prove the main theorem of the section, Theorem 3.7, in §3.3. Although limited in scope, this result covers some very important manifolds of practical interest: $\mathbb{S}^n$ the unit sphere in $\mathbb{R}^{n+1}$, $SO(3)$ the group of rotations in $\mathbb{R}^3$, and $\mathbb{RP}^n$ the real projective space in $\mathbb{R}^{n+1}$. In Section §4, for more general manifolds, we derive two classes of sub-optimal convergence conditions: In §4.1 we give a result (Theorem 4.1) in which convergence is guaranteed at the expense of smaller spread of data points, whereas in §4.2 (Theorem 4.2) the allowable step-size is compromised to guarantee convergence. In §5 we study how (as a result of curvature) for data points having an elongated configuration the convergence can be slow. Finally, in §6, we slightly deviate from the main theme of the paper and study global convergence (i.e., with arbitrary initial condition) of the algorithm. In this case guaranteed convergence to the global center is out of question and the difficulties associated with non-differentiability of the cost function become more visible. Nevertheless, for certain manifolds (e.g., $\mathbb{S}^n$ and $SO(3)$) we show that the constant step-size gradient descent algorithm behaves (more or less) desirably and under generic conditions converges to a *local* center of mass.

2. Preliminaries, a conjecture, and prior work.

2.1. Preliminaries on the Riemannian Center of Mass and the Gradient Descent Algorithm.

2.1.1. Notations and Conventions. Let M be an n-dimensional complete[6] Riemannian manifold with distance function d.[7] In view of the Hopf-Rinow Theorem [42, p. 84], by "complete" more precisely we mean complete and *connected*. We denote the tangent space at $x \in M$ by T_xM and by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ we mean the Riemannian structure and the corresponding norm, respectively (dependence on the base point is implicit and clear from the context). By a C^k function in a subset of M we mean a continuous function which is k^{th} order continuously differentiable in the subset as commonly understood in differential geometry. For a function $f : M \to \mathbb{R}$, ∇f denotes its gradient vector field with respect to $\langle \cdot, \cdot \rangle$. By eigenvalues of the Hessian of f at x we mean the eigenvalues of its Hessian -a bilinear form in T_xM- represented by a symmetric matrix in an orthogonal basis of T_xM.[8] We assume that the sectional curvatures of M are bounded from above and below by Δ and δ, respectively. The exponential map of M at $x \in M$ is denoted by $\exp_x(\cdot) : T_xM \to M$ and its inverse (wherever defined) is denoted by $\exp_x^{-1}(\cdot)$. The injectivity radius of M is denoted by $\mathrm{inj}M$ and we assume $\mathrm{inj}M > 0$. An open ball with center $o \in M$ and radius ρ is

[6] Completeness of the manifold is not necessary for most of our results (since they are local) and our results could be easily adapted to e.g., non-singular regions of a singular manifold. However, for this purpose the statements of our results could become rather cumbersome.

[7] In our definitions relating to Riemannian manifolds we mainly follow [42].

[8] While in (Euclidean) optimization literature "an eigenvalue of the Hessian" is a familiar term, in Riemannian geometry literature this term is not used since the quantity it refers to is not intrinsically defined (because it is only invariant under orthogonal change of coordinates in the tangent space and not more general linear change of coordinates). However, in this paper, considering the possible background of our readers, we prefer to use this term (with careful attention to its exact meaning).

denoted by $B(o,\rho)$ and its closure by $\bar{B}(o,\rho)$. By $X^\top$, $\|X\|_F$, and $\|X\|_2$ for a matrix X we mean its transpose, Frobenius norm, and 2-norm, respectively.

2.1.2. Convex functions and sets in Riemannian manifolds. Convexity plays an important role in our developments and we have the following definition:

Definition 2.1. *Let A be an open subset of M such that every two points in A can be connected by at least one geodesic of M such that this geodesic lies entirely in A. Assume that $f : A \to \mathbb{R}$ is a continuous function. Then f is called (strictly) convex if the composition $f \circ \gamma : [0,1] \to \mathbb{R}$ is (strictly) convex for any geodesic $\gamma : [0,1] \to A$. We say that f is globally (strictly) convex if it is (strictly) convex in M.*

If f is C^2 in A, then convexity (strict convexity) of f is equivalent to $\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\gamma(t))|_{t=0} \geq 0$ (> 0), where $\gamma : [0,1] \to A$ is any geodesic in A.

An insightful fact is the following [51]:

Theorem 2.2. *The only globally convex function in a compact Riemannian manifold is a constant function.*

Therefore, at least on compact manifolds, we have no hope in building globally convex functions. However, if we restrict ourselves to smaller subsets of M, we can build nontrivial convex functions. For that purpose, *strongly convex* subsets of M are particularly suitable, because they are quite similar to standard convex sets in $\mathbb{R}^n$:

Definition 2.3. *A set $A \subset M$ is called strongly convex if any two points in A can be connected by a unique minimizing geodesic in M and the geodesic segment lies entirely in A.*

Define

$$r_{\mathrm{cx}} = \frac{1}{2}\min\{\mathrm{inj}M, \frac{\pi}{\sqrt{\Delta}}\}, \tag{2.1}$$

with the convention that $\frac{1}{\sqrt{\Delta}} = \infty$ for $\Delta \leq 0$. An open ball $B(o,\rho)$ with $\rho \leq r_{\mathrm{cx}}$ is strongly convex in M; the same holds for any closed ball $\bar{B}(o,\rho)$ if $\rho < r_{\mathrm{cx}}$ (see e.g., [42, p. 404] and [14, pp. 168-9]). In fact, $B(o,\rho)$ with $\rho \leq r_{\mathrm{cx}}$ is even more similar to a convex set in Euclidean space: for any $x, y \in B(o,\rho)$ the minimal geodesic from x to y is the only geodesic connecting them which lies entirely in $B(o,\rho)$.

2.1.3. Differentiability and convexity properties of the distance function and estimates on its Hessian. Now, we briefly give some facts about the Riemannian distance function which will be used throughout the paper. Let $y \in M$. The function $x \mapsto d(x,y)$ is continuous for every $x \in M$ and it is differentiable (in fact C^∞) in $M \setminus (\{y\} \cup \mathcal{C}_y)$, where $\mathcal{C}_y$ is the cut locus of y (see e.g., [42, pp. 108-110]). We recall the notion of cut locus: Let $\tilde{D}_y \in T_yM$ be the largest (star-shaped) domain containing the origin of T_yM in which $\exp_y : T_y \to M$ is a diffeomorphism, and let $\tilde{C}_y$ be the boundary of $\tilde{D}_y$. Then $\mathcal{C}_y = \exp_y(\tilde{C}_y)$ is called the *cut locus* of y and $x \in \mathcal{C}_y$ is called a *cut point* of y. One has $M = \exp_y(\tilde{D}_y \cup \tilde{C}_y)$ (see e.g., [14, pp. 117-118] or [42, p. 104]). The distance between y and $\mathcal{C}_y$ is called the *injectivity radius* of y denoted by inj$y$, and by definition $\mathrm{inj}M = \inf_{y\in M} \mathrm{inj}y$. It is well known that $\mathcal{C}_y$ is a closed set in M of measure zero; therefore, the distance function is smooth almost everywhere in M and it is Lipschitz continuous in M. In a general manifold M the differentiability property of $x \mapsto d(x,y)$ at $x = y$ is similar to the case where $M = \mathbb{R}^n$ equipped with the standard Euclidean metric; in particular, $x \mapsto \frac{1}{2}d^2(x,y)$ is a C^∞ function in $M \setminus \mathcal{C}_y$. However, the behavior at far away points (namely, close to the cut locus) is of substantially different nature and depends on the topology and curvature of M (recall that in Euclidean space the cut locus of every point is empty). As a matter of fact, the cut locus of any point contains all information about the topology of M. Understanding the cut locus is difficult in general, but it is known that there are two types of cut points: *ordinary* and *singular* [11]. Let $x \in \mathcal{C}_y$, then x is an ordinary cut point of y if there are at least two minimal geodesics from x to y, otherwise x is called a singular cut point. For example, on the unit sphere $\mathbb{S}^n$ the cut locus of every point is its antipode which is an ordinary cut point. It is known that ordinary cut points of y are dense in $\mathcal{C}_y$. The distance function clearly is *not* differentiable at an ordinary cut point, but in general it can be C^1 (but not C^2) at a singular cut point [11].

Next, we recall some useful estimates about the Hessian of the Riemannian distance function. We adopt the following definitions:

$$\mathrm{sn}_\kappa(l) = \begin{cases} \frac{1}{\sqrt{\kappa}}\sin(\sqrt{\kappa}l) & \kappa > 0 \\ l & \kappa = 0 \\ \frac{1}{\sqrt{|\kappa|}}\sinh(\sqrt{|\kappa|}l) & \kappa < 0 \end{cases} \quad \mathrm{ct}_\kappa(l) = \begin{cases} \sqrt{\kappa}\cot(\sqrt{\kappa}l) & \kappa > 0 \\ \frac{1}{l} & \kappa = 0 \\ \sqrt{|\kappa|}\coth(\sqrt{|\kappa|}l) & \kappa < 0, \end{cases} \tag{2.2}$$

and

$$\mathrm{b}_\kappa(l) = \begin{cases} \sqrt{\kappa}l\cot(\sqrt{\kappa}l) & \kappa \geq 0 \\ 1 & \kappa < 0 \end{cases} \quad \mathrm{c}_\kappa(l) = \begin{cases} 1 & \kappa \geq 0 \\ \sqrt{|\kappa|}l\coth(\sqrt{|\kappa|}l) & \kappa < 0. \end{cases} \tag{2.3}$$

Assume that $x \in M$ (distinct from y) is such that $d(x,y) < \min\{\mathrm{inj}y, \frac{\pi}{\sqrt{\Delta}}\}$.[9] Furthermore, assume that $t \mapsto \gamma(t)$ with $\gamma(0) = x$ is a unit speed geodesic making, at x, an angle β with the *minimal* geodesic from y to x. It can be proved that (see e.g., [42, pp. 152-154])

$$\mathrm{ct}_\Delta(d(x,y))\sin^2\beta \leq \frac{\mathrm{d}^2}{\mathrm{d}t^2}d(\gamma(t),y)\big|_{t=0} \leq \mathrm{ct}_\delta(d(x,y))\sin^2\beta, \tag{2.4}$$

where ct_κ is defined in (2.2). Based on the above one can verify that

$$\mathrm{b}_\Delta(d(x,y)) \leq \frac{\mathrm{d}^2}{\mathrm{d}t^2}\big(\frac{1}{2}d^2(\gamma(t),y)\big)\big|_{t=0} \leq \mathrm{c}_\delta(d(x,y)), \tag{2.5}$$

and more generally that

$$d^{p-2}(x,y)\min\{p-1,\mathrm{b}_\Delta(d(x,y))\} \leq \frac{\mathrm{d}^2}{\mathrm{d}t^2}\big(\frac{1}{p}d^p(\gamma(t),y)\big)\big|_{t=0} \leq d^{p-2}(x,y)\max\{p-1,\mathrm{c}_\delta(d(x,y))\}. \tag{2.6}$$

We will use these two relations very often; in doing so it is useful to have in mind that $\mathrm{b}_\kappa(l)$ and $\mathrm{c}_\kappa(l)$, respectively, are decreasing and increasing in l.

Remark 2.4. We emphasize that although in the left hand side of the above bounds only Δ appears explicitly both the curvature and topology of M determine the convexity properties of the distance function. Specifically, the requirement $d(x,y) < \mathrm{inj}y$ should not be overlooked (in particular, since at a cut point the distance function often becomes non-differentiable). Notice that Δ gives (some) information about the Riemannian curvature tensor of M and $\mathrm{inj}y$ (or $\mathrm{inj}M$) gives (some) information about the topology of M. For example, $\mathbb{R}$ and the unit circle $\mathbb{S}^1$ both have zero sectional curvature, while $\mathrm{inj}\mathbb{S}^1 = \pi$ and the injectivity radius of $\mathbb{R}$ is infinity. Now obviously $x \mapsto \frac{1}{2}d^2(x,y)$ in $\mathbb{R}$ is globally convex, while in $\mathbb{S}^1$ it is not (as seen directly or as a

[9] Instead of this condition, it is often more convenient to require $d(x,y) < \min\{\mathrm{inj}M, \frac{\pi}{\sqrt{\Delta}}\}$, which is a more conservative yet global version of the condition.

consequence of Theorem 2.2). The interesting fact is that $x \mapsto \frac{1}{2}d^2(x,y)$ has positive definite Hessian in $\mathbb{S}^1 \setminus \{y'\}$, where y' is the antipodal point of y, but since at y' it is not differentiable, it is not (globally) convex over $\mathbb{S}^1$.

2.1.4. Riemannian L^p center of mass.

We start by the following definition:

Definition 2.5. *The (global) Riemannian L^p center of mass or mean (a.k.a. Fréchet mean) of the data set $\{x_i\}_{i=1}^N \subset M$ with respect to weights $0 \leq w_i \leq 1$ ($\sum_{i=1}^N w_i = 1$) is defined as the minimizer(s) of*

$$f_p(x) = \begin{cases} \frac{1}{p}\sum_{i=1}^N w_i d^p(x,x_i) & 1 \leq p < \infty \\ \max_i d(x,x_i) & p = \infty, \end{cases} \tag{2.7}$$

in M. We denote the center by $\bar{x}_p$. We call a local minimizer of f_p a local center of mass of the data set with respect to the weights.[10]

The reader is referred to [3] for details and other related definitions. As a convention when referring to the center of mass of some data points we usually do not refer to explicit weights unless needed. As another convention when p is not specified we assume $p = 2$, which is the most commonly used case. Although $p = 1$ and $p = \infty$ are also used often, in this paper our focus is limited to $2 \leq p < \infty$. The reason is that in our analysis we require f_p to be twice-continuously differentiable (in a small region at least) and we determine the constant step-size of the gradient algorithm in terms of the upper bounds on the eigenvalues Hessian of f_p. As in the Euclidean case, in the more general Riemannian case also one can see from (2.6) that for $1 \leq p < 2$ the Hessian of f_p might be unbounded. It is known that Lipschitz continuous gradient (in particular bounded Hessian) is necessary for the convergence of a gradient descent method with constant-step size [41].

Although $f_p : M \to \mathbb{R}$ is a globally convex function when M is a Euclidean space (or more generally an Hadamard manifold) it is not globally convex when M is an arbitrary manifold. In particular, the center of mass might not be unique; however, if the data points are close enough, then the center is unique. The following theorem gives sufficient conditions for existence and uniqueness of the Riemannian center of mass.

Theorem 2.6. *Let $2 < p < \infty$. Consider $\{x_i\}_{i=1}^N \subset B(o,\rho)$ and assume $0 \leq w_i \leq 1$ with $\sum_{i=1}^N w_i = 1$. If $\rho \leq r_{cx}$, then the Riemannian L^p center of mass $\bar{x}_p$ is unique, is inside $B(o,\rho)$, and is the unique zero of the gradient vector field ∇f_p in $\bar{B}(o,\rho)$. Moreover if no data point has weight 1, then $\bar{x}_p$ is a non-degenerate critical point of f_p (i.e., the Hessian of f_p at $\bar{x}_p$ is positive-definite).* [11]

[10] Overall, there is no consensus among authors about the terminology and we find it more convenient to use "local" and "global" Riemannian center of mass as defined here. A local minimizer of f_p in M is sometimes called a Karcher mean, although this definition bears little relation with the way Grove and Karcher [25] or Karcher [28] originally defined what they called the Riemannian center of mass (see also [3] for more details). Given $\{x_i\}_{i=1}^N \subset B(o,\rho)$ with small enough ρ those authors defined the center of mass as a zero of ∇f_2 in $\bar{B}(o,\rho)$ or alternatively as a local minimizer of f_2 in $\bar{B}(o,\rho)$. Notice that since f_2 might not be differentiable at the cut locus of each data point it is not a-priori clear that, in general, a local or global minimizer of f_2 in M should coincide with a zero of ∇f_2. It is known that on the circle $\mathbb{S}^1$ a local minimizer of f_2 always coincides with a zero of its gradient (i.e., it is a smooth critical point of f_2, see e.g., [27, 13] and also our Theorem 6.3).

[11] In Theorem 2.1 in [3], the condition on ρ is stated as $\rho < r_{\mathrm{cx}}$, but since we have a finite number of data points the current version follows immediately. Also from the statement of the theorem, $\bar{x}_p$ is the only zero of ∇f_p in $B(o,\rho)$, but from the proof of the it can be seen that the vector field $-\nabla f_p$ is inward-pointing on the boundary of $B(o,\rho)$, and hence $\bar{x}_p$ is the only zero in the entire $\bar{B}(o,\rho)$. Similarly, the fact about non-degeneracy is not present in the statement of the theorem, however, it is proved in its proof.

For a proof see [3]. Also for $1 \leq p < 2$ and $p = \infty$ see [3] and [50]. We refer the reader to [50, 26, 20] and [7] for algorithms for $p = 1$ and $p = \infty$, respectively.

2.1.5. Gradient descent algorithm for finding the Riemannian center of mass. For later reference we derive the intrinsic gradient descent algorithm for locating the Riemannian L^p center of mass (see [1] or [49] for an introduction to optimization on Riemannian manifolds). One can check that

$$\nabla f_p(x) = -\sum_{i=1}^{N} w_i d^{p-2}(x, x_i) \exp_x^{-1} x_i, \tag{2.8}$$

for any $x \in M$ as long as it is not in the cut locus of any of the data points. In particular, if $\{x_i\}_{i=1}^N \subset B(o, \rho)$, where $\rho < \frac{\mathrm{inj} M}{2}$, then for any $x \in B(o, \rho)$ the above expression is well defined in the classical sense (i.e., it is uniquely defined, cf. §6.2). Notice that the above expression is well defined for almost every $x \in M$, because the set at which f_p is not differentiable has measure zero (for $p > 1$ this set is $\cup_i \mathcal{C}_{x_i}$ and for $p = 1$ it is $\cup_i \mathcal{C}_{x_i} \cup \{x_i\}_i$). As we will see this non-differentiability has severe implications on the behavior of the constant step-size gradient descent.

Algorithm 1 is a gradient descent algorithm for locating the Riemannian L^p center of mass of $\{x_i\}_{i=1}^N$.

```
1. Consider $\{x_i\}_{i=1}^N \subset B(o,\rho) \subset M$ and weights $\{w_i\}_{i=1}^N$
   ($0 \leq w_i \leq 1$, $\sum_i^N w_{=}1$). Choose $x^0 \in M$.
2. if $\nabla f_p(x^k) = 0$ then stop, else set

                 $x^{k+1} = \exp_{x^k}(-t_k \nabla f_p(x^k))$          (2.9)

   where $t_k > 0$ is an “appropriate” step-size and $\nabla f_p(\cdot)$ is
   defined in (2.8).
3. goto step 2.
```

Algorithm 1: Gradient descent for finding the Riemannian L^p center of mass.

Besides practical considerations (e.g., stopping criterion), at least two important issues are left unspecified in Algorithm 1, namely, how to choose x^0 and how to choose t_k for each k. The most natural choice for x^0 is one point in $B(o, \rho)$, say one of the data points (unless when $p = 1$). Note that in practice o and the exact value of ρ might not be known.

The choice of t_k is more complicated. The next general proposition gives a prescription for a step-size interval which ensures reducing the cost function at an iteration of a the algorithm, provided one knows an upper bound on the eigenvalues of the Hessian of the cost function in a region in which the iterates live. The proof of this proposition follows from the second order Taylor series expansion (with remainder).

PROPOSITION 2.7. *Let $x \in M$ and consider an open neighborhood $S \subset M$ containing x. Let $f : M \to \mathbb{R}$ be a function whose restriction to S is twice-continuously differentiable and let the real number H_S be an upper bound on the eigenvalues of the Hessian of f in S. There exists $t_{x,S} > 0$ such that for all $t \in [0, t_{x,S})$ the curve $t \mapsto \exp_x(-t\nabla f(x))$ does not leave S and*

$$f(\exp_x(-t\nabla f(x))) \leq f(x) - \|\nabla f(x)\|^2 t + \frac{H_S \|\nabla f(x)\|^2}{2} t^2. \tag{2.10}$$

For $t \in (0, \min\{t_{x,S}, \frac{2}{H_S}\})$, with the convention that $\frac{1}{H_S} = +\infty$ for $H_S \leq 0$, we have $f(\exp_x(-t\nabla f(x))) \leq f(x)$ with equality only if x is a critical point of f. Moreover, when $H_S > 0$ the right hand side of (2.10) is minimized for $t = \frac{1}{H_S}$.

Notice that the fact that for $t \in [0, t_{x,S})$ the curve $t \mapsto \exp_x(-t\nabla f(x))$ stays in S is crucial in enabling us to use the upper bound H_S and derive (2.10). This concept appears frequently in our analysis and it is useful to have the following definition:

Definition 2.8. *Let $x^k \in S \subset M$. We say that iterate x^{k+1} of Algorithm 1 stays in S if $x^{k+1} = \exp_{x^k}(-t_k\nabla f_p(x^k)) \in S$. We say that the iterate x^{k+1} of Algorithm 1 continuously stays in S if $\exp_{x^k}(-s\nabla f_p(x^k)) \in S$ for $s \in [0, t_k]$.* Obviously, continuously staying in S is stronger than staying in S. However, they are equivalent under some conditions:

Proposition 2.9. *If S is a strongly convex set and $t_k\|\nabla f_p(x^k)\| < \mathrm{inj}M$ for every $k \geq 0$, then for the iterates of Algorithm 1 staying in S implies (and hence is equivalent to) continuously staying in S.*

Proof. Assume that x^k and x^{k+1} both belong to S. Recall that $t \mapsto \exp_x(-t\nabla f_p(x^k))$ for $t \in [0, t_k]$ is a minimizing geodesic if $t_k\|\nabla f_p(x^k)\| < \mathrm{inj}M$. Therefore, by strong convexity of S, $t \mapsto \exp_x(-t\nabla f_p(x^k))$ for $t \in [0, t_k]$ must be the only minimizing geodesic between x^k and x^{k+1} and must lie in S entirely. □

The following convergence result is a standard one when the cost is C^2 (or at least has Lipschitz gradient) and is globally convex; however, our version is adapted to f_p which, in general, is neither globally C^2 nor convex. The assumption of the theorem that each iterate of the algorithm continuously stays in a neighborhood S of $\bar{x}_p$, in which $\bar{x}_p$ is the only zero of ∇f_p, is a crucial enabling ingredient of the proof. In fact, our goal in §3 and §4 is essentially to identify such a neighborhood (under certain conditions).

Theorem 2.10. *Let $2 \leq p < \infty$ and assume that $\bar{x}_p$ is the center of mass of $\{x_i\}_{i=1}^N \subset B(o, \rho)$, where $\rho \leq r_{cx}$. Let S be a bounded open neighborhood of $\bar{x}_p$ such that f_p is C^2 in S and C^1 in $\bar{S}$, the closure of S. Furthermore, assume that $\bar{x}_p$ is the only zero of the vector field ∇f_p in $\bar{S}$. Let H_S be an upper bound on the eigenvalues of the Hessian of f_p in S. In Algorithm 1 choose $t_k = t \in (0, \frac{2}{H_S})$. If starting from $x^0 \in S$, each iterate of Algorithm 1 continuously stays in S, then $f_p(x^{k+1}) \leq f_p(x^k)$ for $k \geq 0$ with equality only if $x^k = \bar{x}_p$, and x^k converges to $\bar{x}_p$.*

Proof. Since by assumption $x^k \in S$ for $k \geq 0$ and $\bar{S}$ is compact, there is a subsequence $\langle x^{k_j}\rangle_{k_j}$ converging to a point $x^* \in \bar{S}$. By Proposition 2.7 we have $f_p(x^{k+1}) \leq f_p(x^k)$ with equality only if $x^k = \bar{x}_p$, and furthermore

$$t(1 - \frac{H_S t}{2}) \sum_{j=1}^{k} \|\nabla f_p(x^j)\|^2 \leq f_p(x^{k+1}) - f_p(x^0), \tag{2.11}$$

for every $k \geq 0$. Since $\langle f_p(x^k)\rangle_k$ is a bounded sequence, the above implies that $\|\nabla f_p(x^k)\| \to 0$; hence, by continuity of ∇f_p we have $\|\nabla f_p(x^*)\| = 0$, that is, x^* is a zero of ∇f_p in $\bar{S}$. But by the assumption about S this means that x^* coincides with $\bar{x}_p$ and therefore $x^{k_j} \to \bar{x}_p$. Notice that by the same argument any infinite subsequence of $\langle x^k\rangle_k$ has a subsequence which converges to $\bar{x}_p$. But that is enough to complete the proof, because if $\langle x^k\rangle_k$ does not converge to $\bar{x}_p$, there must be an $\epsilon > 0$ and an infinite subsequence of $\langle x^k\rangle_k$ which stays away from $B(\bar{x}_p, \epsilon)$, and that cannot happen. □

Next, we give a very simple but insightful example.

*Example 2.11 (*Finding the mean of two points on the unit circle $\mathbb{S}^1$*).* Let M be the unit circle $\mathbb{S}^1$ centered at the origin $(0, 0)$ and equipped with the standard arc length distance d. Recall that $\Delta = 0$ and $\mathrm{inj}M = \pi$. Let o denote the point $(1,0)$ (see Figure 2.1). We consider two data points $x_1, x_2 \in \mathbb{S}^1$ represented as $x_i = (\cos\theta_i, \sin\theta_i)$ $(i = 1,2)$ where $0 < \theta_1 < \rho \leq \frac{\pi}{2}$ and $\theta_2 = -\theta_1$. We specify the weights and θ_1 later. Under the mentioned assumption that $\rho \leq \frac{\pi}{2}$, Theorem 2.6 guarantees that the center of mass $\bar{x}$ is unique and in fact one can see that $\bar{x} = (\cos\bar{\theta}, \sin\bar{\theta})$ where $\bar{\theta} = w_1\theta_1 + w_2\theta_2$. More importantly, it also follows that $\bar{x}$ is the unique zero of ∇f_2 in $B(o,\rho)$ (as well as $B(o,\frac{\pi}{2})$). Notice that f_2 is smooth within $B(o, \pi - \theta_1)$ (it does not follow from Theorem 2.6 but it is an easily verifiable fact that $\bar{x}$ is the unique zero of ∇f_2 in $B(o, \pi - \theta_1)$). On the other hand, f_2 is not differentiable at x'_1 and x'_2, the antipodal points of x_1 and x_2, respectively. Furthermore, in $\mathbb{S}^1 \setminus \{x'_1, x'_2\}$ the Hessian of f_2 is defined and is equal to 1, hence the largest possible range of the constant step-size $t_k = t$ is the interval $(0,2)$. Next, we see under what conditions Theorem 2.10 applies. It is easy to check that, independent of the weights, with $x^0 \in B(o,\rho)$ and step-size $t_k = t \in (0,1]$ the iterates of Algorithm 1 continuously stay in $B(o,\rho)$. Therefore, an acceptable S is the ball $B(o,\rho)$ and Theorem 2.10 ensures convergence to the global center $\bar{x}$ provided step-size t is in the interval $(0,1]$. This result is essentially not different from what we have in $\mathbb{R}$. However, the situation for $t \in (1,2)$ is rather subtle since with a large step-size the iterates might leave the ball $B(o,\rho)$ or even $B(o,\pi-\theta_1)$ and enter a region in which there is another zero of ∇f_2. To be specific notice that f_2 can be parameterized with $\theta \in (-\pi, +\pi]$ as

$$f_2(\theta) = \frac{1}{2}\begin{cases} w_1(\theta - \theta_1 - 2\pi)^2 + w_2(\theta-\theta_2)^2 & -\pi < \theta \leq \theta_1 - \pi \\ w_1(\theta-\theta_1)^2 + w_2(\theta-\theta_2)^2 & \theta_1 - \pi \leq \theta \leq \theta_2 + \pi \\ w_1(\theta-\theta_1)^2 + w_2(\theta - \theta_2 + 2\pi)^2 & \theta_2 + \pi \leq \theta \leq \pi. \end{cases} \tag{2.12}$$

Now, let us fix $\theta_1 = \frac{2\pi}{5}$ (and $\theta_2 = -\frac{2\pi}{5}$). First, let $w_1 = \frac{1}{10}$ and $w_2 = \frac{9}{10}$. The solid curve in the right panel in Figure 2.1 shows the graph of $f_2(\theta)$. The two cranks in the curve at $\theta'_1 = \frac{-3\pi}{5}$ and $\theta'_2 = \frac{3\pi}{5}$ are due to the non-differentiability of f_2 at antipodal points of x_1 and x_2. If we run Algorithm 1 with $x^0 = x_1$ and step-size $t = \frac{25}{18}$ we have $x^1 = x'_1$, thus x^1 coincides with a non-differentiable critical point of f_2 at which the algorithm is, in fact, not well defined. In the generic setting the probability of this happening is zero; however, for larger t, x^1 will leave $B(o, \pi - \theta_1)$. It can be seen that for this specific pair of weights ∇f_2 has only one zero in $\mathbb{S}^1$. Consequently, in practice, Algorithm 1 for almost every initial condition in $\mathbb{S}^1$ and step-size $t_k = t$ in the interval $(0,2)$ will find the global center of mass $\bar{x} = (\cos\bar{\theta}, \sin\bar{\theta})$, where $\bar{\theta} = \frac{-8\pi}{25}$ (this fact does not follow from Theorem 2.10 but is not difficult to verify in this special case, see also §6 and Corollary 6.7). But we might not be this lucky always! For example, let $w_1 = \frac{1}{4}$ and $w_2 = \frac{3}{4}$. The dashed curve in Figure 2.1 shows $f_2(\theta)$ for this pair of weights. One can verify that in addition to the global minimizer $\bar{\theta} = \frac{-\pi}{5}$, this time, $f_2(\theta)$ has a local minimizer (which is not global) at $\bar{\theta}' = \frac{-7\pi}{10}$. Now if we run Algorithm 1 with $x^0 = x_1$ and constant step-size $t_k = t$, where $t = \frac{11}{6}$, then we have $x^1 = \frac{-7\pi}{10}$, i.e., the next iterate coincides with the local center $\bar{x}' = (\cos\bar{\theta}', \sin\bar{\theta}')$ and the algorithm gets stuck at the wrong center! For values of t slightly smaller or larger than $\frac{11}{6}$ the algorithm still converges to $\bar{x}'$. Notice that this happens despite the fact that the cost is reduced at each iteration.[12] This simple example only shows the difficulties stemming from the topology of a manifold and not from its curvature.

[12] It would be interesting to see whether an example exists in which due to the non-differentiability of f_2, we have $f_2(x^1) > f_2(x^0)$ if x^1 does not continuously stay in S. Such a phenomenon could lead to an oscillatory behavior (see §6). In §6.2 we show that in $\mathbb{S}^1$ and certain other manifolds this situation cannot happen.

Nevertheless, it should be clear that in order for Algorithm 1 to have a predictable or desirable behavior which is as data-independent as possible it is important to identify conditions under which the assumptions of Theorem 2.10 are satisfied (mainly that the iterates continuously stay in a candidate S).

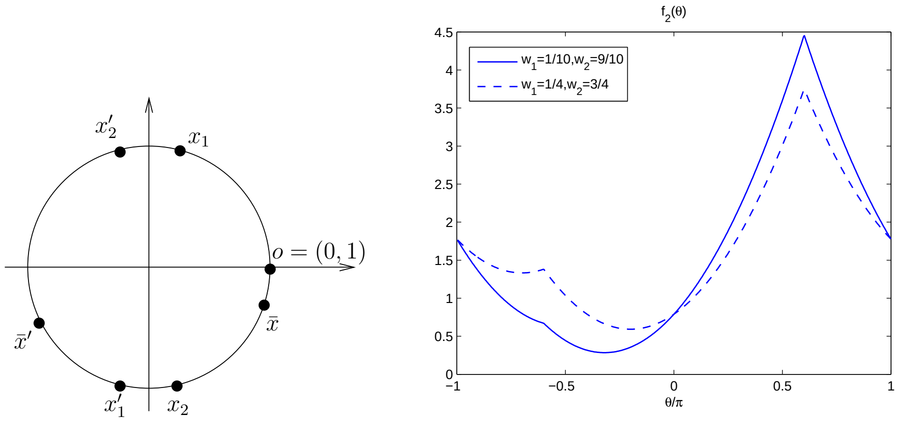

Fig. 2.1: The left panel shows the configuration of data points x_1 and x_2 in Example 2.11 and the right panel is the graph of $f_2(\theta)$ for two different pairs of weights. The function f_2 is non-differentiable at x_1' and x_2', the antipodal points of x_1 and x_2. The example shows that if the step-size is large the iterates might leave the region in which f_2 is smooth and the algorithm might converge to $\bar{x}'$, a local center of x_1 and x_2, instead of the global center $\bar{x}$ (despite the fact that the cost is reduced at each step). Notice that the correct way of thinking about the plotted graphs it to visualize them while identifying points $\theta = -\pi$ and $\theta = +\pi$ or to think of them as periodic graphs with period 2π.

2.1.6. Speed of convergence and the best step-size. In Proposition 2.7, $t = \frac{1}{H_S}$ is the best step-size in the sense that in each iteration it causes the largest decrease in the upper bound of $f_p(x^{k+1})$ described in the right hand side of (2.10). The following theorem relates this choice to the speed of convergence of the algorithm. The proof of the theorem is adopted from [49, p. 266, Theorem 4.2] where a proof is given for a globally convex C^2 function. Here, we adapt that proof to constant step-size gradient descent for minimizing f_p (which is only locally C^2 and convex).

Theorem 2.12. *Let $2 \leq p < \infty$. Let $\bar{x}_p$ be the L^p center of mass of $\{x_i\}_{i=1}^N \subset B(o,\rho) \subset M$, where $\rho \leq r_{cx}$. Suppose that S is a strongly convex neighborhood around $\bar{x}_p$ in which f_p is twice-continuously differentiable, and let h_S and H_S, respectively, denote a lower and upper bound on the eigenvalues of the Hessian of f_p in S. Furthermore, assume that S is small enough such that one can choose $h_S > 0$. In Algorithm 1 choose a constant step-size $t_k = t \in (0, \frac{2}{H_S})$. If after a finite number of iterations k' each iterate continuously stays in S, then for $k \geq k'$ we have*

$$d(x^k, \bar{x}_p) \leq K q^{\frac{k-k'}{2}}. \tag{2.13}$$

In the above K and q are defined as

$$K = \left(\frac{2(f_p(x^{k'}) - f_p(\bar{x}_p))}{h_S}\right)^{\frac{1}{2}} \text{ and } q = 1 - \alpha(1 - \frac{\alpha}{2})\frac{h_S}{H_S}(1 + \frac{h_S}{H_S}), \tag{2.14}$$

where $\alpha = tH_S$. In particular, $0 \leq q < 1$ and $x^k \to \bar{x}_p$ as $k \to \infty$.

Proof. Let $\gamma(t) = \exp_x(t \exp_x^{-1} \bar{x}_p)$ be the minimal geodesic from $x \in S$ to $\bar{x}_p$. Note that $\gamma(t) \in S$ for $t \in [0,1]$ due to strong convexity of S. After writing the second order Taylor's series of $t \mapsto f_p(\gamma(t))$ around $t = 0$ and using the bounds on the Hessian of f_p one gets

$$-d(x,\bar{x}_p)\|\nabla f_p(x)\| + \frac{h_S}{2}d^2(x,\bar{x}_p) \leq f_p(\bar{x}_p) - f_p(x) \leq d(x,\bar{x}_p)\|\nabla f_p(x)\| + \frac{H_S}{2}d^2(x,\bar{x}_p). \tag{2.15}$$

Similarly by expansion of $t \mapsto f_p(\gamma(t))$ around $t = 1$ one gets

$$\frac{h_S}{2}d^2(x,\bar{x}_p) \leq f_p(x) - f_p(\bar{x}_p) \leq \frac{H_S}{2}d^2(x,\bar{x}_p) \tag{2.16}$$

Also notice that by the first order Taylor series expansion of $t \mapsto \nabla f_p(\gamma(t))$ around $t = 1$ we have

$$h_S d(x,\bar{x}_p) \leq \|\nabla f_p(x)\| \leq H_S d(x,\bar{x}_p) \tag{2.17}$$

Next, we plug the lower bound on $d^2(x,\bar{x}_p)$, from the right inequality in (2.16), into the left inequality in (2.15). Hence, (after reordering) we have

$$f_p(x) - f_p(\bar{x}_p) \leq d(x,\bar{x}_p)\|\nabla f_p(x)\| - \frac{h_S}{H_S}(f_p(x) - f_p(\bar{x}_p)). \tag{2.18}$$

Combining this with the left inequality in (2.17) results in

$$h_S(1 + \frac{h_S}{H_S})\big(f_p(x) - f_p(\bar{x}_p)\big) \leq \|\nabla f_p(x)\|^2. \tag{2.19}$$

Now assume $k \geq k'$ so $x^k, x^{k+1} \in S$. Then subtracting $f_p(\bar{x}_p)$ from both sides of (2.10) and using (2.19) both at $x = x^k$ yield

$$f_p(x^{k+1}) - f_p(\bar{x}_p) \leq q\big(f_p(x^k) - f_p(\bar{x}_p)\big). \tag{2.20}$$

Therefore, we have

$$f_p(x^k) - f_p(\bar{x}_p) \leq \big(f_p(x^{k'}) - f_p(\bar{x}_p)\big)q^{k-k'}, \tag{2.21}$$

for $k \geq k'$. Now combining this with the left inequality in (2.16) yields (2.13). □

This theorem predicts a *lower bound* on the speed of convergence (i.e., the actual convergence is not worse than what the theorem predicts). The accuracy of this prediction, in part, depends on the accuracy of our estimates of the lower and upper bounds on the eigenvalues of the Hessian. Observe that when we are only given H_S, $\alpha = 1$ (i.e., $t_k = \frac{1}{H_S}$) gives the smallest a-priori q. We call $t_k = \frac{1}{H_S}$ the *best a-priori* step-size given H_S.

Remark 2.13 (Asymptotic q). Notice that a strongly convex S which works in Theorem 2.10 might not work in Theorem 2.12 (since the assumption $h_S > 0$ might not hold true) and a smaller S might be needed for this theorem. Nevertheless, if we start with an S (and a corresponding H_S) for which Theorem 2.10 holds, then there exists a smaller strongly convex $S'(\subset S)$ for which $h_{S'} > 0$ and Theorem 2.12 holds. H_S is still an upper bound on the eigenvalues of the Hessian of f_p in S' and in lack of any knowledge about S' still $t = \frac{1}{H_S}$ is the best a-priori step-size. The actual asymptotic speed of convergence is determined by q in a very small neighborhood S' of $\bar{x}_p$. In fact, in the limit $h_{S'}$ and $H_{S'}$ converge, respectively, to $\lambda'_{\min}$ and $\lambda'_{\max}$ the smallest and largest eigenvalues of the Hessian of f_p at $\bar{x}_p$. Therefore, for any step-size $t \in (0, \frac{2}{\lambda'_{max}})$, we define the associated *asymptotic* q, denoted by q', where in (2.14), h_S, H_S, and α are replaced, respectively, by $\lambda'_{\min}$, $\lambda'_{\max}$, and $\alpha' = t\lambda'_{\max}$. Notice that if $t = \frac{1}{H_S}$, then $\alpha' \leq 1$ and the smaller the H_S the smaller the q' will be. In general, "smaller H_S" means that either we make S smaller or we choose a more accurate upper bound on the eigenvalues of the Hessian of f_p in S.

2.1.7. Relation to contraction mapping. Some parts of convergence analysis in Theorem 2.12 and in particular (2.13) are reminiscent of the behavior of contraction mapping iterations. Ideally one would like to have the map $F_p : S \to M$ defined by $F_p(x) = \exp_x(-t\nabla f_p(x))$ to be a contraction mapping. Recall that a map $F : S \to M$ is a contraction mapping on $S \subset M$ if for all $x, y \in S$ we have $d(F(x), F(y)) \leq \kappa d(x, y)$ with $\kappa < 1$ and $F(S) \subset S$ (i.e., F preserves S). Then F has a unique fixed point in S to which the iterates $x^{k+1} = F(x^k)$ converge with $x^0 \in S$. In particular, we have $d(F(x^{k+1}), \bar{x}) \leq \kappa d(x^k, \bar{x})$ and a similar convergence rate as (2.13) would result. It is interesting to note that from (2.20) and (2.16) we get

$$d(x^{k+1}, \bar{x}_p) \leq \sqrt{q\frac{H_S}{h_S}} d(x^k, \bar{x}_p). \tag{2.22}$$

Note that $\sqrt{q\frac{H_S}{h_S}}$ is not necessarily smaller than 1. However, the next result shows that ultimately, when the iterates get close enough to a non-degenerate local minimizer, then the gradient descent iteration acts as a contraction mapping.

Proposition 2.14 (Asymptotic stability of a non-degenerate local minimizer via contraction mapping). *Let $\bar{x}$ be a non-degenerate local minimizer of $f : M \to \mathbb{R}$. Assume that f is C^2 is an open ball around $\bar{x}$. Let $t \in (0, \frac{2}{H_{\bar{x}}})$, where $H_{\bar{x}}$ is the largest eigenvalue of the Hessian of f at $\bar{x}$. Then for small enough r the map $F(x) = \exp_x(-t\nabla f(x))$ is a contraction mapping in $B(\bar{x}, r)$ with unique fixed point $\bar{x}$. In particular, a gradient descent algorithm with step-size t and starting in $B(\bar{x}, r)$ will converge to $\bar{x}$.*

Proof. Let us denote the derivative of $F : M \to M$ at x by F_{*x}. Now in an orthonormal basis in $T_{\bar{x}}M$ we have $F_{*\bar{x}} = I_n - t\text{Hess}f$, where Hess$f$ is the matrix representation of the Hessian of f (this relation is intuitively obvious and its exact proof follows e.g., from derivations in [23, §2]). Now for the 2-norm of Hess$f$ we have $\kappa = \|F_{*x}\|_2 = |1 - th_{\bar{x}}| < 1$, where $h_{\bar{x}} > 0$ is the smallest eigenvalue of Hess$f$. But since f is C^2 (and hence F_{*x} is continuous in x), we see that for small enough r we can find κ' where $\kappa \leq \kappa' < 1$ such that $\|F_{*x}\|_2 \leq \kappa' < 1$ for all $x \in B(\bar{x}, r)$ (we can make r small enough to make sure that the ball is strongly convex too). Now let $\gamma : [0, 1] \to M$ be the minimal geodesic from x_* to $x \in B$, then $F(\gamma(s))$ is a curve from $\bar{x}$ to $F(x)$ whose speed is not larger than that of γ at very $s \in [0, 1]$, hence length$(F(\gamma([0, s]))) \leq$ length$(\gamma([0, s]))$ for $s \in [0, 1]$. This implies that $d(\bar{x}, F(x)) \leq d(\bar{x}, x)$ and $F(x) \in B$ for $x \in B(\bar{x}, r)$. Consequently, now, if $\eta : [0, 1] \to M$ is the minimal geodesic from $x \in B$ to $y \in B$, we have $F(\eta(s)) \in B$, and we see that length$(F(\eta([0, s]))) \leq \kappa'$length$(\eta([0, s]))$. This implies that $d(F(x), F(y)) \leq \kappa' d(x, y)$. □

The above proof would have been considerably more difficult if we wanted to give explicit (sharp) bounds on r or if we wanted to show that F is a contraction mapping in a ball with a center different from the fixed point of F (see [33, 23] and our discussion in §2.3).

2.2. A Conjecture: The best convergence condition. As mentioned before, reducing the cost at each iteration is not enough to guarantee the convergence of Algorithm 1 to the global center of mass. Nevertheless, we conjecture that if the constant step-size is chosen not too large and the initial condition is not far from $\bar{x}_p$ (as specified next), then the cost f_p can be reduced at each iteration, the iterates stay close to $\bar{x}_p$ and converge to it .

Conjecture 2.15. *Let $p = 2$ and assume that $\bar{x}_2$ is the L^2 center of mass of $\{x_i\}_{i=1}^N \subset B(o,\rho) \subset M$ where $\rho \leq r_{\text{cx}}$. Let $H_{B(o,\rho)} = \mathrm{c}_\delta(2\rho)$, where c_κ is defined in (2.3). In Algorithm 1, assume $x^0 \in B(o,\rho)$ and choose a constant step-size $t_k = t$, for some $t \in (0, \frac{1}{H_{B(o,\rho)}}]$. Then we have the following: Each iterate continuously stays in $B(o,\rho)$ (and hence the algorithm will be well defined for every $k \geq 0$), $f_2(x^{k+1}) \leq f_2(x^k)$ $(k \geq 0)$ with equality only if $x^k = \bar{x}_2$, and $x^k \to \bar{x}_2$ as $k \to \infty$. More generally, for $2 \leq p < \infty$ the same results hold if $t \in (0, \frac{1}{H_{B(o,\rho),p}}]$, where $H_{B(o,\rho),p} = (2\rho)^{p-2}\max\{p-1, \mathrm{c}_\delta(2\rho)\}$.*

Now we explain the sense in which this conjecture is the best result one can hope for. We narrow down our desired class of convergence conditions to a class which gives conditions that are uniform in the data sets and in the initial condition. More specifically, we consider the following general and natural class of conditions:

> Convergence Condition Class (C): Consider Algorithm 1 and fix $2 \leq p < \infty$, and let δ and Δ, respectively, be a lower and upper bound on sectional curvatures of M. Specify the largest $\bar{\rho}$ $(0 < \bar{\rho} \leq r_{\text{cx}})$ such that for every $\rho \leq \bar{\rho}$ there are
>
> 1. a number $\rho'_{\delta,\Delta,\rho,p}$ $(\rho \leq \rho'_{\delta,\Delta,\rho,p} \leq r_{\text{cx}})$ depending only on δ, Δ, ρ, and p; and
> 2. another number $t_{\delta,\Delta,\rho,\rho',p}$ depending only on δ, Δ, ρ, p, and $\rho'_{\delta,\Delta,\rho,p}$,
>
> for which the following holds: for every ball $B(o,\rho) \subset M$, for every set of data points in $B(o,\rho)$, for every set of weights in (2.7), for every initial condition in $B(o,\rho)$, and for every constant step-size $t_k = t \in (0, t_{\delta,\Delta,\rho,\rho',p}]$, each iterate of Algorithm 1 continuously stays in $B(o,\rho'_{\delta,\Delta,\rho,p})$, and $x^k \to \bar{x}_p$.

First, notice that with $\bar{\rho} > r_{\text{cx}}$ there is no hope to have convergence to the global center (in this class), since the global center might not lie in $B(o,\bar{\rho})$ and in general ∇f_p might have more than one zero in $B(o,\bar{\rho})$. Next, observe that Conjecture 2.15 belongs to this class of conditions and it claims that $\bar{\rho} = r_{\text{cx}}$ is achievable; therefore, in this sense the conjecture claims the best possible condition in this class. In particular, this means that the conjecture gives the best possible spread of data points and the largest region of convergence, i.e., it allows both $\{x_i\}_{i=1}^N \subset B(o,r_{\text{cx}})$ and $x^0 \in B(o,r_{\text{cx}})$.

Now let us see in what sense the step-size interval in Conjecture 2.15 is optimal. One can verify that $H_{B(o,\rho),p}$ in Conjecture 2.15 is the smallest *uniform* upper bound on the eigenvalues of the Hessian of f_p in $B(o,\rho)$. Here, by a uniform bound we mean a bound which is independent of the data points, the weights, and o. Based on Remark 2.13, if we know that each iterate continuously stays in $B(o,\rho'_{\delta,\Delta,\rho,p})$ and further if we only know $H_{B(o,\rho'),p}$, then from Theorem 2.12 we see that $t_k = \frac{1}{H_{B(o,\rho'),p}}$ is the best uniform a-priori step-size (in the sense that it yields the smallest uniform a-priori q). Next, notice that, with this t_k, the smaller the ρ', the smaller the $H_{B(o,\rho'),p}$ and hence the smaller the q' (the asymptotic q) will be, see Remark 2.13. Consequently, $t_k = \frac{1}{H_{B(o,\rho'),p}}$ at $\rho' = \rho$ gives the smallest *asymptotic* q among all *best uniform a-priori* step-sizes $t_k = \frac{1}{H_{B(o,\rho'),p}}$, where $\rho \leq \rho' \leq r_{\text{cx}}$. Conjecture 2.15 claims that, indeed, $\rho'_{\delta,\Delta,\rho,p}$ can be as small as ρ (independent of δ, Δ, p). Therefore, in summary, among all conditions in class (C), the sub-class which prescribes $t_{\delta,\Delta,\rho,\rho',p} = \frac{1}{H_{B(o,\rho'_{\delta,\Delta,\rho,p}),p}}$ allows to have the smallest uniform a-priori q (by choosing $t_k = \frac{1}{H_{B(o,\rho'_{\delta,\Delta,\rho,p}),p}}$), and in this sub-class, Conjecture 2.15 gives the largest $t_{\delta,\Delta,\rho,\rho',p}$, therefore it allows for the largest step-size and hence the *smallest uniform asymptotic* q in this sub-class. We stress that this sense of optimality of the step-size interval should not be constructed as giving the best speed of convergence for any actual data configuration; rather it gives the best uniform lower bound on the speed of convergence in Theorem 2.12. This means that for all data configurations, weights, and initial conditions in $B(o,\rho)$ the actual speed of convergence will not be worse than the one predicted by Theorem 2.12 where the associated q in (2.14) is the best uniform a-priori q. Quite similarly, one could argue that the step-size interval in Conjecture 2.15 is optimal in the sense that it allows for the largest decrease per iteration in the upper bound given in (2.10) of Proposition 2.7.

The proof of the conjecture in the case of manifolds with zero curvature is quite easy and straightforward. However, the general case seems to be difficult. The main difficulty in proving Conjecture 2.15 is in proving that the iterates continuously stay in $B(o,\rho)$. Nevertheless, in Theorem 3.7 we prove the conjecture for manifolds of constant nonnegative curvature and 2-dimensional manifolds of variable nonnegative curvature. As this proof suggests, we believe that the difficulty in proving this conjecture has more to do with geometry (than optimization) and the need of good estimates (which currently seem not to exist) about the behavior of the exponential map in a manifold. Our proof of Theorem 3.7 certainly constitutes some strong evidence that the conjecture also is true for more manifolds of nonnegative curvature. For manifolds with negative curvature we have also some evidence that the conjecture is true. For example, the conjecture is trivially true if all the data points are concentrated at a single point in $B(o,\rho)$, and by continuity, it is also true if all the data points are concentrated enough around a single point in $B(o,\rho)$. Weaker convergence results can be established with some efforts. For example, in §4, we derive weaker convergence results in Theorems 4.1 and 4.2. As a comparison, Theorem 4.1 gives smaller allowable spread and smaller region of convergence than Conjecture 2.15. Theorem 4.2, on the other hand, gives allowable spread and region of convergence which could be very close to $B(o,r_{\text{cx}})$, but the allowable step-size is restricted significantly in this theorem.

Remark 2.16 (Related to Remark 2.13). It is useful to put this conjecture in some context, especially in view of Theorems 2.10 and 2.12 and Remark 2.13. It should be clear from our discussions in §2.1.4 and Remark 2.13, that when $\Delta > 0$, the conjecture claims convergence for initial conditions in regions in which the Hessian of f_p is not necessarily positive-definite. Therefore, what really could help us in proving this result is Theorem 2.10 and not Theorem 2.12. Although, already assured of convergence, we can use Theorem 2.12 to give us an asymptotic behavior of the algorithm. All our proved convergence theorems (which are Theorems 3.7, 4.1, and 4.2) are proved using Theorem 2.10. In Theorems 3.7 and 4.1 both the initial conditions and the trajectories of the algorithm can lie in regions in which only this theorem applies. The case of Theorem 4.1 is rather interesting. Under the conditions of Theorem 4.1

the initial condition must lie in a region in which f_p happens to be strictly convex (since $\frac{1}{3}r_{\rm cx} < \frac{\pi}{4\sqrt{\Delta}}$, see §2.1.4), however, the trajectory of the algorithm can visit a region in which the Hessian of f_p is not positive-definite.

2.3. Prior work. There are not many *accurate* and *correctly proven* [13]results available about the convergence of gradient descent (and more specifically constant-step-size) for locating the Riemannian center of mass. For constant step-size algorithms, the most accurate and useful results are due to Le [33] and Groisser [23, 24] for the L^2 mean. In view of Proposition 2.14, we put Le's and Groisser's approaches in further context. Given $\{x_i\}_{i=1}^N \subset B(o,\rho)$, Le gives an explicit bound on ρ such that $F_2(x) = \exp_x(-\nabla f_2(x))$ is a contraction mapping in $B(\bar{x}_2,\rho)$, hence she proves that with $x^0 = o$ the iterates converge to $\bar{x}_2$. Groisser on the other hand gives an explicit bound on ρ such that F_2 is a contraction mapping in $B(o,\rho)$, hence he proves the convergence of the algorithm and also gives a constructive proof for the uniqueness of the center of mass. Groisser's results are more general and allow to analyze both Newton's method and the gradient descent method, while Le's result gives a slightly better bound on the allowable spread of the data points. In particular, Le shows that when M is a locally symmetric manifold of nonnegative sectional curvature and $\rho \leq \frac{3}{10}r_{\rm cx}$, F_2 is a contraction mapping when restricted to $B(\bar{x}_2,\rho)$ provided $x^0 = o$. Since $\bar{x}_2$ is a fixed point of F_2, *starting* from o the iterates will not leave $B(\bar{x}_2,\rho)$. Le's result leaves room for significant improvement in the allowable spread of data points compared to our Conjecture 2.15. Notice that Le's result can be used to deduce convergence for an *arbitrary* initial condition in $B(o,\rho)$ assuming a ρ half as before, that is $\rho \leq \frac{3}{20}r_{\rm cx}$. This is obviously a more practical scenario. On the other hand, both Theorem 2.10 and Theorem 2.12 show that gradient descent can converge for initial conditions outside the (relatively small) region in which it acts as a contraction mapping. In particular, by using Theorem 2.10 we manage to prove larger domains of convergence than those provided by Le and Groisser. However, it should be noted in such larger domains, in general, the algorithm might be slow since initially no contraction property exists and moreover with larger data spread the asymptotic q might be smaller (since the the convexity of the cost function at $\bar{x}_2$ might be less). Our Theorem 4.1 (which needs only $\rho \leq \frac{1}{3}r_{\rm cx}$ and does not require local symmetry or nonnegative curvature) is a considerable improvement over Le's result. Still our Theorem 3.7, which is the best one can hope for in the case of manifolds of constant nonnegative curvature, is an even further improvement over Le's result (when applied to these manifolds).

In [34] a convergence result is given for a (hard-to-implement) gradient method which varies the step-size in order to confine the iterates to a small ball. A result in [31] is somewhat similar in nature to our Theorem 4.1, yet it does not yield an explicit convergence condition. Local convergence of Algorithm 1 with $t_k = 1$ on $\mathbb{S}^n$ under the generic condition of "x^0 being close enough to the center" is argued in [12]; however, such a condition is of little practical use. A few authors have also studied other related problems and methods e.g., stochastic gradient methods [5], projected gradient methods [32], Newton's method [12, 23], and variable step-size gradient algorithm for the L^1 mean or median [50, 26]. We add that distance based definition is not the only

[13] A mistake made by some authors (see e.g., [35] and [20]) in proving such results has been to wrongly assume that $f_p : M \to \mathbb{R}$ is *globally* convex (or strictly convex) if the data points are in a small enough ball, which in general is not true, e.g., if M is compact (see Theorem 2.2). In particular, in [35] or [20] no effort has been made to show that the iterates continuously stay in a region in which the global center of mass is the only local minimizer of f_p.

way to define averages and other authors also have considered special group theoretic or algebraic structures in defining averages, see e.g., [19, 18, 39, 22] or [16, Ch. 20].

3. Convergence on Manifolds of Constant Nonnegative Curvature and 2-Dimensional Manifolds of Nonnegative Curvature (An Optimal Result).

In this section, we prove Theorem 3.7 which is essentially Conjecture 2.15 for the spacial case of a 2-dimensional manifold of nonnegative curvature or a manifold of constant nonnegative curvature.

3.1. A useful triangle secant comparison result.

Here, we prove a comparison result used to prove Theorem 3.7 (see Figure 3.1 and Theorem 3.4).

Theorem 3.1. *Let M be a 2-dimensional manifold of nonnegative curvature or an n-dimensional manifold of constant nonnegative curvature ($n \geq 2$). Let $x, y_1, y_2 \in M$ be three points that lie in a ball of radius r_{cx} defined in (2.1). Assume that the internal angle $\angle y_1 x y_2$ is equal to α ($0 < \alpha \leq \pi$). Consider another triangle in $\mathbb{R}^2$ (or $\mathbb{R}^n$) with vertices $\tilde{x}, \tilde{y}_1, \tilde{y}_2$ and assume that the internal angles at x and $\tilde{x}$ and their corresponding sides are equal in the two triangles. Consider a geodesic γ in M passing through x and making angles $\alpha_1 > 0$ and $\alpha_2 > 0$ ($\alpha_1 + \alpha_2 = \alpha$) with minimal geodesic sides xy_1 and xy_2, respectively. Denote by m the point where γ meets the minimal geodesic side $y_1 y_2$ for the first time. Similarly, let a secant line of triangle $\tilde{y}_1 \tilde{x} \tilde{y}_2$ passing through $\tilde{x}$ make angles α_1 and α_2 with sides $\tilde{x}\tilde{y}_1$ and $\tilde{x}\tilde{y}_2$, respectively, and denote by $\tilde{m}$ the point where this secant line meets the side $\tilde{y}_1\tilde{y}_2$. Then the secant segment xm is not smaller than the secant segment $\tilde{x}\tilde{m}$, moreover, it is longer than the secant segment $\tilde{x}\tilde{m}$ if M is of constant positive curvature.*

Before proving the theorem, we remark that this comparison is only meaningful when M is either 2-dimensional or of constant curvature, because otherwise there is no guarantee that the geodesic from x would meet the opposite geodesic side. In the case of a constant curvature manifold M the enabling property is the well-known *axiom of plane*: Let $x \in M$ and assume that $W \subset T_x M$ is a k-dimensional subspace of $T_x M$, then the set $\exp_x(W \cap B(0_x, \rho))$ is a totally geodesic submanifold of M. Here $B(0_x, \rho)$ is the open ball of radius ρ around the origin of $T_x M$ and $0 < \rho < \text{inj}M$ (see e.g., [42, p. 136]). In a 2-dimensional manifold the geodesic from x meeting the opposite is obvious. However, in both cases, in fact, we need an extra size condition which ensures that the geodesics are unique and that is why the assumption on x, y_1, y_2 being inside a ball of radius r_{cx} is made.

Now, we prove the theorem. We give two proofs: The first one is a direct one using direct computation and applies only to the constant curvature case, the second one is indirect but more general and uses Toponogov's comparison theorem. This proof is mostly due to Marc Arnaudon [4].

Proof. First proof (sketch, only for constant nonnegative curvature): The case of constant zero curvature is obvious. We prove the theorem for $\mathbb{S}^2$ (where $\Delta = 1$ and $\text{inj}M = \pi$), the more general case of constant positive curvature follows immediately (especially by including the extra size restriction due to finite injectivity radius of M). Let us denote the lengths of minimal geodesic sides xy_1, xy_2, y_1y_2 by b, c, and a respectively. Denote the length of the geodesic secant segment xm by $z(b, c; \alpha_1, \alpha_2)$. Using spherical trigonometric identities (e.g., [36, p. 53]) one can show that (see also [36, p. 55])

$$\cot z(b, c; \alpha_1, \alpha_2) = \frac{\cot b \sin \alpha_2 + \cot c \sin \alpha_1}{\sin(\alpha_1 + \alpha_2)}. \tag{3.1}$$

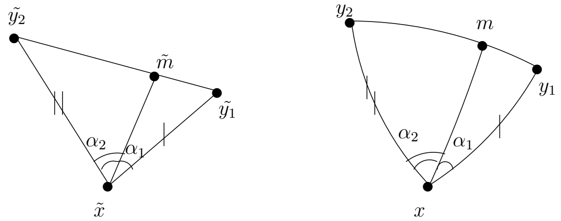

Fig. 3.1: $\triangle y_1xy_2$ is a triangle in a manifold of constant positive curvature (or a 2-dimensional manifold of nonnegative curvature) and $\triangle\tilde{y}_1\tilde{x}\tilde{y}_2$ is the corresponding triangle in Euclidean space. Corresponding equal angles and sides are marked. According to Theorem 3.1, the geodesic secant xm is longer than the secant $\tilde{x}\tilde{m}$.

Similarly, denote the length of the secant segment $\tilde{x}\tilde{m}$ by $\tilde{z}(b,c,\alpha_1,\alpha_2)$. It is easy to see that

$$\tilde{z}(b,c;\alpha_1,\alpha_2)=\frac{bc\ \sin(\alpha_1+\alpha_2)}{b\sin\alpha_1+c\sin\alpha_2}, \tag{3.2}$$

where in both relations $\alpha_1+\alpha_2=\alpha$. Note that z and $\tilde{z}$ are both smaller than π and therefore to show $z(b,c;\alpha_1,\alpha_2)>\tilde{z}(b,c;\alpha_1,\alpha_2)$ we could show $\cot z(b,c;\alpha_1,\alpha_2)<\cot\tilde{z}(b,c;\alpha_1,\alpha_2)$ with $\alpha_1+\alpha_2=\alpha$ (see (3.1) and (3.2)). The result then, essentially, follows from strict concavity of $t\mapsto g(t)=\cot\frac{1}{t}$ in the interval $(\frac{1}{\pi},\infty)$.

Second Proof [4]: We first prove the following angle comparison result [4]:

Lemma 3.2. *We have $\angle xy_1y_2\geq\angle\tilde{x}\tilde{y}_1\tilde{y}_2$ or $\angle xy_2y_1\geq\angle\tilde{x}\tilde{y}_2\tilde{y}_1$ with strict inequality in the case of positive constant curvature.*

Proof. This can be proved using the triangle version of Toponogov's comparison theorem with lower curvature bound (see e.g., [40, p. 337-8]). It is more convenient to use Figure 3.2 for this purpose. In this figure the first triangle is $\triangle y_1xy_2$ in M and the second triangle is the auxiliary triangle $\triangle\tilde{y}_1'\tilde{x}\tilde{y}_2$ in $\mathbb{R}^2$ whose sides are correspondingly equal to those of $\triangle y_1xy_2$ (corresponding equal sides and angles are marked). From Toponogov's theorem we have $\angle y_1xy_2\geq\angle\tilde{y}_1'\tilde{x}\tilde{y}_2'$, $\angle xy_1y_2\geq\angle\tilde{x}\tilde{y}_1'\tilde{y}_2'$, and $\angle xy_2y_1\geq\angle\tilde{x}\tilde{y}_2'\tilde{y}_1'$ with strict inequality in the constant curvature case with $\Delta>0$. Now comparing $\triangle\tilde{y}_1\tilde{x}\tilde{y}_2$ with $\triangle\tilde{y}_1'\tilde{x}\tilde{y}_2'$ (both in $\mathbb{R}^2$), we see that since $\angle\tilde{y}_1\tilde{x}\tilde{y}_2=\alpha\geq\angle\tilde{y}_1'\tilde{x}\tilde{y}_2'$, we must have $\angle\tilde{x}\tilde{y}_1'\tilde{y}_2'\geq\angle\tilde{x}\tilde{y}_1\tilde{y}_2$ or $\angle\tilde{x}\tilde{y}_2'\tilde{y}_1'\geq\angle\tilde{x}\tilde{y}_2\tilde{y}_1$; and therefore, $\angle xy_1y_2\geq\angle\tilde{x}\tilde{y}_1\tilde{y}_2$ or $\angle xy_2y_1\geq\angle\tilde{x}\tilde{y}_2\tilde{y}_1$, with strict inequality in the constant positive curvature case. □

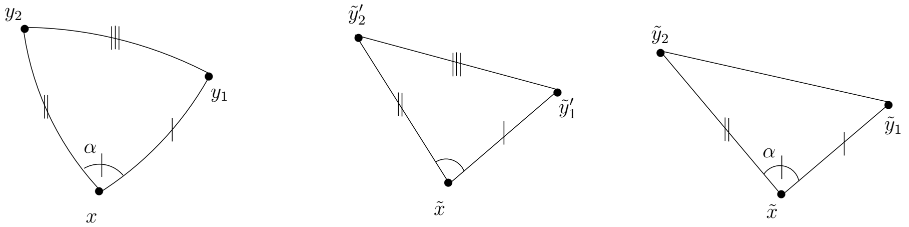

Fig. 3.2: Lemma 3.2 is proved using Toponogov's triangle comparison theorem via the auxiliary triangle $\triangle\tilde{x}\tilde{y}_1'\tilde{y}_2'$.

Next, we use an infinitesimal argument. Let us assume $\angle xy_1y_2\geq\angle\tilde{x}\tilde{y}_1\tilde{y}_2$. Consider the minimal geodesic side $\eta:[0,\alpha]\to M$ (resp. $\tilde{\eta}:[0,\alpha]\to M$) from y_1 to y_2

(resp. $\tilde{y}_1$ to $\tilde{y}_2$) with $\eta(0) = y_1$ and $\eta(\alpha) = y_2$ (resp. $\tilde{\eta}(0) = \tilde{y}_1$ and $\tilde{\eta}(\alpha) = \tilde{y}_2$). Denote the distance functions in M and $\mathbb{R}^2$, by d and $\tilde{d}$, respectively. The claim is equivalent to $d(\eta(t), x) \geq \tilde{d}(\tilde{\eta}(t), \tilde{x})$ for every $t \in (0, \alpha)$ with strict inequality in the constant positive curvature case. This is clearly true for small enough t, because $\angle xy_1y_2 \geq \angle\tilde{x}\tilde{y}_1\tilde{y}_2$. Let t_0 be the first time after which this relation is violated. Then we must have

$$d(\eta(t_0), x) = \tilde{d}(\tilde{\eta}(t_0), \tilde{x}) \text{ and } \angle x\eta(t_0)y_2 < \angle\tilde{x}\tilde{\eta}(t_0)\tilde{y}_2. \tag{3.3}$$

Now by applying Lemma 3.2 to the triangles $\triangle x\eta(t_0)y_2$ and $\triangle\tilde{x}\tilde{\eta}(t_0)\tilde{y}_2$, we must have $\angle xy_2y_1 \geq \angle\tilde{x}\tilde{y}_2\tilde{y}_1$, which requires $d(\eta(t)), x) \geq \tilde{d}(\tilde{\eta}(t), \tilde{x})$ for t close to α. Then by continuity, this implies that for some t_1 $((t_0 < t_1 < \alpha))$

$$d(\eta(t_1)), x) = \tilde{d}(\tilde{\eta}(t_1), \tilde{x}) \text{ and } \angle x\eta(t_1)\eta(t_0) < \angle\tilde{x}\tilde{\eta}(t_1)\tilde{\eta}(t_0). \tag{3.4}$$

But (3.3) and (3.4) in the triangles $\triangle x\eta(t_0)\eta(t_1)$ and $\triangle\tilde{x}\tilde{\eta}(t_0)\tilde{\eta}(t_1)$ contradict Lemma 3.2, hence $\angle x\eta(t_0)y_2 \geq \angle\tilde{x}\tilde{\eta}(t_0)\tilde{y}_2$ and the claim must hold, at least, with non-strict inequality. However, a careful look at the proof reveals that the same proof can be used to prove the claim for the positive constant curvature case. □

3.2. Riemannian pointed convex combinations. Intuitively, one would like to think of $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ as a Riemannian convex combination with similar properties as the Euclidean convex combination. We call this a *Riemannian pointed convex combination* of $\{x_i\}_{i=1}^N \subset M$ with respect to $x \in M$ and with weights $(w_1, \ldots, w_N)$ $(\sum_{i=1}^N w_i = 1)$. [14] For the case of a manifold of constant nonnegative curvature or a 2-dimensional manifold of nonnegative curvature we show that this is a valid definition. The following general proposition is useful in making sense of Riemannian pointed convex combinations as well as proving Theorem 3.7.

Proposition 3.3. *Let $S \subset M$ be a strongly convex set containing $\{x_i\}_{i=1}^N$ and let $x \in S$. Assume that for arbitrary weights $0 \leq w_1, w_2 \leq 1$ (with $w_1 + w_2 = 1$) and for any $y_1, y_2 \in S$, $\exp_x\left(t(w_1 \exp_x^{-1} y_1 + w_2 \exp_x^{-1} y_2)\right) \in S$ for $t \in [0, 1]$. Then for every set of points $\{x_i\}_{i=1}^N \subset S$ and corresponding weights $0 \leq w_i \leq 1$ (with $\sum_i w_i = 1$), $\exp_x(t\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ also belongs to S for $t \in [0, 1]$.*

Proof. We prove the claim for $N = 3$ and for larger N it follows by induction. Let $y(t) = \exp_x(t\sum_{i=1}^3 w_i \exp_x^{-1} x_i)$. Note that we can write

$$y(t) = \exp_x\left(t\Big(w_1 \exp_x^{-1} x_1 + (1 - w_1)\big(\frac{w_2}{w_2 + w_3} \exp_x^{-1} x_2 + \frac{w_3}{w_2 + w_3} \exp_x^{-1} x_3\big)\Big)\right). \tag{3.5}$$

Since by assumption $\tilde{x}_2 = \exp_x(\frac{w_2}{w_2+w_3} \exp_x^{-1} x_2 + \frac{w_3}{w_2+w_3} \exp_x^{-1} x_3)$ belongs to S, there exists a unique minimizing geodesic between x and $\tilde{x}_2$. Therefore, $\exp_x^{-1} \tilde{x}_2$ is well defined and belongs to the injectivity domain of $\exp_x$ and we have $\exp_x^{-1} \tilde{x}_2 = \frac{w_2}{w_2+w_3} \exp_x^{-1} x_2 + \frac{w_3}{w_2+w_3} \exp_x^{-1} x_3$. Since $\exp_x\left(t(w_1 \exp_x^{-1} x_1 + (1 - w_1) \exp_x^{-1} \tilde{x}_2)\right)$ belongs to S (by our assumption) we must have $y(t) \in S$ for $t \in [0, 1]$. □

An example of such a set S is the convex hull of $\{x_i\}_{i=1}^N$. Recall that the convex hull of $A \subset M$ (if it exists) is defined as the smallest strongly convex set containing

[14] Notice that if $M = \mathbb{R}^n$, $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ translates to $x + \sum_{i=1}^N w_i(x_i - x)$, which is independent of x. In a nonlinear space the pointed convex combination does not enjoy this base-point independence, and that is why we have been explicit in calling $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ a pointed convex combination with respect to x. Moreover, to have "nice" properties, x cannot be arbitrary and must belong to the convex hull of $\{x_i\}_{i=1}^N$, as explained in Remark 3.5.

A. If A lies in a strongly convex set obviously its convex hull exists. It is known that the convex hull of a finite set of points in a constant curvature manifold is a closed set. Also in a manifold of constant curvature the L^p center of mass ($1 < p < \infty$) of $\{x_i\}_{i=1}^N$ with weights $w_i \geq 0$ belongs to the convex hull of $\{x_i\}_{i=1}^N$ and if $w_i > 0$ for every i, it belongs to the interior of the hull [3]. The next theorem describes the relation between the Riemannian pointed convex combination and the convex hull:

Theorem 3.4. *Let M be a Riemannian manifold of constant non-negative curvature or a 2-dimensional manifold of nonnegative curvature. Let S be a strongly convex set containing $\{x_i\}_{i=1}^N$. Assume that S lies in a ball of radius of at most r_{cx}. For every $x \in S$ and arbitrary weights $w_i \geq 0$ (with $\sum_{i=1}^N w_i = 1$), we have $\exp_x(t \sum_{i=1}^N w_i \exp_x^{-1} x_i) \in S$ for $t \in [0,1]$. In particular, if S is the convex hull of $\{x_i\}_{i=1}^N$, then the convex combination $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ belongs to S for every $x \in S$.*

Proof. By Proposition 3.3, it suffices to show that for arbitrary $y_1, y_2 \in S$ and weights (w_1, w_2) with $w_1 + w_2 = 1$ we have $\tilde{\tilde{m}}(t) = \exp_x \big(t(w_1 \exp_x^{-1} y_1 + w_2 \exp_x^{-1} y_2)\big) \in S$ for $\in [0,1]$. This follows from comparison Theorem 3.1. To see this, first note that, in triangle $\triangle xy_1y_2$ in Figure 3.1, there is a $1-1$ correspondence between the weight pairs (w_1, w_2) and the angle pairs (α_1, α_2), where $\alpha_1 + \alpha_2 = \alpha$. If $x, y_1, y_2 \in S$, then by strong convexity of S we have $m \in S$. Since the distance between x and $\tilde{\tilde{m}}(1)$ is nothing but the length of secant $\tilde{x}\tilde{m}$ in triangle $\triangle \tilde{x}\tilde{y}_1\tilde{y}_2$, it follows from Theorem 3.1 that $\tilde{\tilde{m}}(t)$ must belong to S for $t \in [0,1]$. □

Remark 3.5. Notice that it follows from Theorem 3.1 that, if x is not in the convex hull of $\{x_i\}_{i=1}^N$, then $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ does not, necessarily, belong to the convex hull. Hence, having x in the convex hull is necessary in the above proposition. We also mention that Buss and Fillmore, define the notion of (spherical) Riemannian convex combination, as the Riemannian center of mass of the data points $\bar{x}_2$, when the weights w_i's vary [12]. We have used the term "pointed convex combination" to distinguish our definition from Buss and Fillmore's. Buss and Fillmore show that, in $\mathbb{S}^n$, the convex combination defined in this fashion fills the convex hull of the data points as the weights are varied. However, from Theorem 3.1 we see that the pointed convex combination (while x is fixed) does not enjoy this property. In that sense the pointed convex combination is a weak "convex combination."

Remark 3.6. One might wonder in what directions the above theorem can be extended. One can show that in a manifold of constant *negative* curvature or in a 2-dimensional manifold of negative curvature the inequality in Theorem 3.1 holds in the reverse direction, that is, the secant in the manifold is shorter than the corresponding secant in $\mathbb{R}^n$. This by itself implies that, in a manifold of constant negative curvature, $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ *does not*, necessarily, belong to the convex hull of $\{x_i\}_{i=1}^N$; and one needs to scale down the tangent vector $\sum_{i=1}^N w_i \exp_x^{-1} x_i$ to ensure that it belongs to the convex hull. This scaling somehow should be related to the size of the convex hull or the minimal ball of $\{x_i\}_{i=1}^N$ (see Conjecture 2.15 and Remark 3.8). Recall that the *minimal ball* of $\{x_i\}_{i=1}^N \subset M$ is a closed ball of minimum radius containing $\{x_i\}_{i=1}^N$ (see [3] for more on the minimal ball). Furthermore, even for nonnegative variable curvature in dimension higher than 2, $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ belonging to the convex hull of $\{x_i\}_{i=1}^N$ seems implausible. However, in this case, we conjecture that $\exp_x(\sum_{i=1}^N w_i \exp_x^{-1} x_i)$ belongs to the minimal ball of $\{x_i\}_{i=1}^N$.

3.3. Convergence result. We are now ready to state and prove the main theorem of the section.

THEOREM 3.7. *Assume that M is either a manifold of constant nonnegative curvature $\Delta \geq 0$ or a 2-dimensional manifold with nonnegative curvature upper bounded by $\Delta \geq 0$. Let $p = 2$ and $\{x_i\}_{i=1}^N \subset B(o,\rho)$, where $\rho \leq r_{\rm cx}$ (see (2.1)). In Algorithm 1, choose an initial point $x^0 \in B(o,\rho)$ and a constant step-size $t_k = t$, where $t \in (0,1]$. Then we have: The algorithm is well-defined for every $k \geq 0$, each iterate continuously stays in $B(o,\rho)$, $f_2(x^{k+1}) \leq f_2(x^k)$ with equality only if $x_k = \bar{x}_2$, and $x^k \to \bar{x}_2$ as $k \to \infty$. Moreover, if for some $k' \geq 0$, $x^{k'}$ belongs to the convex hull of $\{x_i\}_{i=1}^N$, then x^k also belongs to the convex hull for $k \geq k'$. More generally, for $2 \leq p < \infty$ the same results hold if we take $t \in (0, t_{\rho,p}]$ with $t_{\rho,p} = \frac{1}{H_{B(o,\rho),p}}$ where $H_{B(o,\rho),p} = (p-1)\big(2\rho\big)^{p-2}$.*

Proof. The fact that each iterate continuously stays in $B(o,\rho)$ follows from Theorem 3.4. The same argument shows that if $x^{k'}$ is in the convex hull of $\{x_i\}_{i=1}^N$, then x^k also belongs to the hull for $k \geq k'$. By Proposition 2.7, step-size $t_k = t$ at each step results in strict reduction of f_2 unless at $\bar{x}_2$. Next, the iterates converging to $\bar{x}_2$ follows from Theorem 2.10 by taking S as $B(o,\rho)$ or the convex hull of $\{x_i\}_{i=1}^N$. For the general p, we notice that $\frac{-1}{H_{B(o,\rho),p}}\nabla f_p(x)$ can be written as $\sum_{i=1}^N \tilde{w}_i \exp_x^{-1} x_i$, where $\sum_{i=1}^N \tilde{w}_i \leq 1$ and $\tilde{w}_i \geq 0$. Therefore, again we can use Theorem 3.4, and the rest of the claims follow similarly. ☐

Remark 3.8. In [23], Groisser introduced the notion of tethering: A map $\Psi : M \to M$ is called *tethered* to $\{x_i\}_{i=1}^N$ if for every strongly convex regular geodesic ball B containing $\{x_i\}_{i=1}^N$, Ψ is defined on B and $\Psi(B) \subset B$. To avoid technical difficulties which probably have little to do with the essence of the property of tethering, we replace "every strongly convex regular geodesic ball" with "every ball of radius less than or equal to $r_{\rm cx}$." Groisser's definition is more general than ours. Groisser conjectured that tethering "might occur fairly generally." Several results in [23] can be strengthened if the tethering assumption holds (even in this weaker sense). In the above theorem, we proved that for $t \in [0,1]$, the map $x \mapsto \exp_x(-t\nabla f_2(x))$ is tethered to $\{x_i\}_{i=1}^N$ in manifolds of constant nonnegative curvature or 2-dimensional manifolds of nonnegative curvature. We conjecture that the same holds for higher dimensional manifolds of nonnegative variable curvature. However, based on the discussion in Remark 3.6, we conjecture that tethering in manifolds of negative curvature does *not* hold. As mentioned in Remark 3.6 (and also expressed in Conjecture 2.15), we conjecture that in order for $x \mapsto \exp_x(-t\nabla f_2(x))$ to map $B(o,\rho) \supset \{x_i\}_{i=1}^N$ to itself, t should be smaller than 1. More specifically, we conjecture that t cannot be independent of ρ, and $t \in [0, \frac{1}{{\rm c}_\delta(2\rho)}]$ suffices.

4. Convergence Results for More General Manifolds.

Here, we prove two classes of results which are sub-optimal compared to Conjecture 2.15. In the first class the spread of data points is compromised to guarantee convergence. In the second class, the step-size is restricted more than the optimal one to ensure that the iterates do not leave a neighborhood in which $\bar{x}_2$ is the only zero of ∇f_2.

4.1. Compromising the spread of data points.

The following theorem is based on the the simple observation that f_p takes larger values outside of $B(o,3\rho)$ than inside of $B(o,\rho)$.

THEOREM 4.1. *Let $p = 2$ and assume that $\bar{x}_2$ is the L^2 center of mass of $\{x_i\}_{i=1}^N \subset B(o,\rho) \subset M$, where $\rho \leq \frac{1}{3} r_{\rm cx}$. Define $t_{\delta,\rho} = \frac{1}{H_{B(o,3\rho)}}$, where $H_{B(o,3\rho)} = {\rm c}_\delta(4\rho)$ and ${\rm c}_\kappa$ is defined in (2.3). In Algorithm 1 assume that $x^0 \in B(o,\rho)$ and for every $k \geq 0$ choose $t_k = t$, where $t \in (0, 2t_{\delta,\rho})$. Then we have the following: The algo-*

rithm is well-defined for all $k \geq 0$ and each iterate of the algorithm continuously stays in $B(o,3\rho)$, $f_2(x^{k+1}) \leq f_2(x^k)$ for $k \geq 0$ (with equality only if $x^k = \bar{x}_2$), and $x^k \to \bar{x}_2$ as $k \to \infty$. Moreover, if x^0 coincides with o, then $\rho \leq \frac{1}{2} r_{\mathrm{cx}}$ is enough to guarantee the convergence, in which case each iterate of the algorithm continuously stays in $B(o,2\rho)$ and we can take $t_{\delta,\rho} = \frac{1}{H_{B(o,2\rho)}}$ where $H_{B(o,2\rho)} = \mathrm{c}_\delta(3\rho)$. More generally, for $2 \leq p < \infty$ the same results hold if we replace $H_{B(o,3\rho)}$ and $H_{B(o,2\rho)}$, respectively, with $H_{B(o,3\rho),p} = (4\rho)^{p-2}\max\{p-1, \mathrm{c}_\delta(4\rho)\}$ and $H_{B(o,2\rho),p} = (3\rho)^{p-2}\max\{p-1, \mathrm{c}_\delta(3\rho)\}$.

Proof. For any $x \in M \setminus B(o,3\rho)$ we have $f_2(x) > 2\rho^2 > f_2(x^0)$ (see (2.7)). From (2.5) and (2.3) and that $\{x_i\}_{i=1}^N \subset B(o,\rho)$, one sees that $H_{B(o,3\rho)} = \mathrm{c}_\delta(4\rho)$ is an upper bound on the eigenvalues of the Hessian of f_2 in $B(o,3\rho)$. Moreover, by Proposition 2.7, for small enough $t \in (0, 2t_{\delta,\rho})$, $s \mapsto \exp_{x^0}(-s\nabla f_2(x^0))$ does not leave $B(o,3\rho)$ for $s \in [0,t]$, and we have $f(\exp_{x^0}(-t\nabla f_2(x^0))) \leq f(x^0)$, with equality only if x^0 is the unique zero of ∇f_2 in $B(o,3\rho)$. However, $s \mapsto \exp_{x^0}(-s\nabla f_2(x^0))$ must lie in $B(o,\rho)$ for all s in $(0, 2t_{\delta,\rho})$, since on the boundary of $B(o,3\rho)$, f is larger than $f(x^0)$ and by continuity $s \mapsto \exp_{x^0}(-s\nabla f_2(x^0))$ cannot leave $B(o,3\rho)$ without making $f_2\big(\exp_{x^0}(-s\nabla f_2(x^0))\big)$ larger than $f_2(x^0)$ inside $B(o,3\rho)$, which is a contradiction. Therefore, for any $t \in (0, 2t_{\delta,\rho})$, the iterate $x^1 = \exp_{x^0}(-t\nabla f_2(x^0))$ continuously stays in $B(o,3\rho)$ and $f_2(x^1) \leq f_2(x^0)$, with equality only if $x^0 = \bar{x}_2$. A similar argument shows that for any $y \in B(o,3\rho)$ such that $f_2(y) \leq f_2(x^0)$, $f(\exp_y(-s\nabla f_2(y)))$ for $s \in [0,t]$ belongs to $B(o,3\rho)$ and $f(\exp_y(-t\nabla f_2(y))) \leq f(y)$ with equality only if $y = \bar{x}_2$. In particular, assuming $x^k \in B(o,3\rho)$ and $f_2(x^k) \leq f_2(x^0)$, by setting $y = x^k \in B(o,3\rho)$, we conclude that x^{k+1} continuously stays in $B(o,3\rho)$ and $f(x^{k+1}) \leq f(x^k)$ with equality only if $\bar{x}^k = \bar{x}_2$. Note that for any point y in $B(o,3\rho) \setminus B(o,\rho)$ we have $d(y,x_i) < 4\rho < \frac{2}{3}\mathrm{inj}M$ for $1 \leq i \leq n$; therefore, $\nabla f_2(y)$ in (2.8) and hence (2.9) in Algorithm 1 are well-defined. Next, by taking $B(o,3\rho)$ as S in Theorem 2.10, we conclude that $x^k \to \bar{x}_2$ as $k \to \infty$. To see the claim about $B(o,2\rho)$, note that if x^0 coincides with o, then we have $f_2(x) > \frac{1}{2}\rho^2 > f_2(x^0)$ for any x out of $B(o,2\rho)$ and the derived conclusions hold with $B(o,2\rho)$. The claims about $2 \leq p < \infty$ follow similarly by further using (2.6). □

4.2. Compromising the step-size. Here, given ρ and ρ' where $\rho < \rho' \leq r_{\mathrm{cx}}$ and assuming the data points lie in $B(o,\rho)$, by restricting the step-size we want to make sure that, starting from $B(o,\rho')$, the iterates do not leave the larger ball $B(o,\rho')$. A rather similar idea has been used in [5] and [50], and here we partially follow the methodology in [50].

For x inside $B(o,\rho)$, let $t_x > 0$ denote the first time $t \mapsto \gamma_x(t) = \exp_x(-t\nabla f_2(x))$ hits the boundary of $B(o,\rho')$. Note that $\sup_{B(o,\rho)} \|\nabla f_2(x)\| < 2\rho$, therefore we must have $t_x > t_x^{\mathrm{in}}$ where

$$t_x^{\mathrm{in}} = \frac{\inf_{x\in B(o,\rho), y \in M\setminus B(o,\rho')} d(x,y)}{2\rho} = \frac{\rho'-\rho}{2\rho}. \tag{4.1}$$

Similarly, for y in the annular region between $B(o,\rho')$ and $B(o,\rho)$, let $t_y > 0$ denote the first time $t \mapsto \gamma_y(t) = \exp_y(-t\nabla f_2(y))$ hits the boundary of $B(o,\rho')$. For $t \mapsto \frac{1}{2}d^2(o,\gamma_y(t))$ one writes the second order Taylor's series expansion in the interval $[0,t_y]$ as:

$$\frac{1}{2}d^2(o,\gamma_y(t_y)) = \frac{1}{2}\rho'^2 = \frac{1}{2}d^2(o,y) + \langle -\nabla f_2(y), -\exp_y^{-1} o\rangle t_y + \frac{1}{2}\frac{\mathrm{d}^2 f_{2,o}(t)}{\mathrm{d}t^2}\Big|_{t=s} t_y^2, \tag{4.2}$$

where s is in the interval $(0,t_y)$. Next, using (2.5) and noting that $\rho^2 - d^2(o,y) > 0$

we verify that

$$t_y > \frac{2\langle -\nabla f_2(y), \exp_y^{-1} o\rangle}{\mathrm{c}_\delta(\rho')}, \tag{4.3}$$

where c_δ is defined in (2.3). Denote by $\angle x_i y o$ the angle, at y, between the minimal geodesics from y to x_i and from y to o. It is shown in Lemma 10 in [50] that

$$\cos\angle x_i y o \geq \frac{\mathrm{sn}_\Delta(d(y,o)-\rho)}{\mathrm{sn}_\Delta(d(y,o)+\rho)}, \tag{4.4}$$

where sn_Δ is defined in (2.2). Using this and observing that $\|\nabla f_2(y)\| \geq d(y,o)-\rho$ we have $t_y > t_y^{\mathrm{out},1}$, where

$$t_y^{\mathrm{out},1} = \frac{2}{\mathrm{c}_\delta(\rho')} \times d(y,o) \times \big(d(y,o)-\rho\big) \times \frac{\mathrm{sn}_\Delta(d(y,o)-\rho)}{\mathrm{sn}_\Delta(d(y,o)+\rho)}. \tag{4.5}$$

Also observe that (trivially) we must have $t_y > t_y^{\mathrm{out},2}$, where

$$t_y^{\mathrm{out},2} = \frac{\rho' - d(y,o)}{\rho + d(y,o)}. \tag{4.6}$$

Obviously, t_y must satisfy $t_y > \max\{t_y^{\mathrm{out},1}, t_y^{\mathrm{out},2}\}$. Define

$$t_{\mathrm{exit}} = \min\{t^{\mathrm{in}}, \inf_{y:\rho\leq d(y,o)<\rho'} \max\{t_y^{\mathrm{out},1}, t_y^{\mathrm{out},2}\}\} \tag{4.7}$$

where t^{in}, $t_y^{\mathrm{out},1}$, and $t_y^{\mathrm{out},2}$ are defined in (4.1), (4.5), and (4.6), respectively, with the assumption $\rho < \rho' \leq r_{\mathrm{cx}}$. We see that for any $z \in B(o,\rho')$ and any $t \in [0, t_{\mathrm{exit}}]$, $\exp_z(-t\nabla f_2(z))$ belongs to $B(o,\rho')$. Notice that $t = t_{\mathrm{exit}}$ is indeed acceptable. Also observe that t_{exit} is larger than zero; since otherwise it can be zero only if for z in the region $B(o,\rho') \setminus B(o,\rho)$ and very close to the boundaries of the region $t_z^{\mathrm{out},1}$ and $t_z^{\mathrm{out},2}$ both become arbitrary close to zero, which obviously cannot happen. Based on this analysis we have the following theorem.

Theorem 4.2. *Let $p = 2$, $\{x_i\}_{i=1}^N \subset B(o,\rho)$ and assume $\rho < \rho' \leq r_{\mathrm{cx}}$. Define $H_{B(o,\rho')} = \mathrm{c}_\delta(\rho'+\rho)$ and set*

$$t^*_{\delta,\Delta,\rho,\rho'} = \min\{t_{\mathrm{exit}}, \frac{1}{H_{B(o,\rho')}}\}, \tag{4.8}$$

where t_{exit} is defined in (4.7). In Algorithm 1, choose an initial condition $x^0 \in B(o,\rho)$ [15] *and step-size $t_k = t$, where $t \in (0, 2t^*_{\delta,\Delta,\rho,\rho'}) \cap [0, t_{\mathrm{exit}}]$. Then we have the following: The algorithm is well-defined for every $k \geq 0$, each iterate continuously stays in $B(o,\rho')$, $f_2(x^{k+1}) \leq f_2(x^k)$ with equality only if $x^k = \bar{x}_2$, and $x^k \to \bar{x}_2$ as $k \to \infty$.*

Proof. The fact that each iterate continuously stays in $B(o,\rho')$ follows from preceding arguments. From this it we see that $d(x^k, x_i) < \mathrm{inj}M$ for every $k \geq 0$ and $1 \leq i \leq N$, and hence the algorithm is well-defined for $k \geq 0$. The rest of the claims follow from Theorem 2.10. □

Next, we give some numerical examples about the interplay between ρ, ρ', and the step-sizes according to Theorem 4.2 and compare that with step-size and allowable

[15] In fact, according to the derivations, one could choose $x^0 \in B(o,\rho')$.

spread from Conjecture 2.15 and Theorem 4.1. First, let $\delta = 0$ and $\Delta > 0$ and let $\rho' = r_{\mathrm{cx}}$. To have $t_k = 1$ we need to have $\rho \leq r_1 \approx 0.0303 r_{\mathrm{cx}}$, while Theorem 4.1 gives much larger ρ, i.e., $\rho \leq \frac{1}{3} r_{\mathrm{cx}}$. We can increase ρ and further restrict the step-size: If we set $\rho = \frac{1}{3} r_{\mathrm{cx}}$, then we get $t^*_{\delta,\Delta,\rho,\rho'} \approx 0.3965$, if $\rho = \frac{9}{10} r_{\mathrm{cx}}$ we get $t^*_{\delta,\Delta,\rho,\rho'} = 0.0353$, and finally when $\rho = 0.99 r_{\mathrm{cx}}$ we get $t^*_{\delta,\Delta,\rho,\rho'} = 0.0033$, all of which are considerably smaller than the optimal step-size of 1 in Conjecture 2.15. Yet, the added value is that we have convergence for more spread-out data points (i.e., going from $\rho \leq \frac{1}{3} r_{\mathrm{cx}}$ to *almost* $\rho \leq r_{\mathrm{cx}}$). Next, let $\delta < 0$, $\Delta = 0$, and $\rho' = \frac{\pi}{2}\sqrt{-\delta}$ (this is just an arbitrary number). To get the optimal step-size in Theorem 2.12 which is $\frac{1}{H_{B(o,\rho')}}$ and is equal to $\frac{1}{\mathrm{c}_\delta(\rho+\rho')}$, we need $\rho \leq r_2 \approx 0.1950\rho'$. Therefore, the $t^*_{\delta,\Delta,\rho,\rho'}$ from Theorem 4.2 cannot be larger than the $t_{\delta,\rho}$ from Theorem 4.1. In fact, if we set $\rho = \frac{1}{3}\rho'$, then we need $t^*_{\delta,\Delta,\rho,\rho'} = 0.3022$ according to Theorem 4.2, while we have $t_{\delta,\rho} = 0.4632$ from Theorem 4.1.

Finer analysis could yield a larger estimate for the exit time than (4.7). However, since $\mathrm{c}_\delta(2\rho)$ is an upper bound on the eigenvalues of the Hessian of f_2 in $B(o,\rho)$, such an improvement will not result in an optimal step-size better than $t_k = \mathrm{ct}_\delta(2\rho)^{-1}$ (cf. (4.8) and Conjecture 2.15).

5. On the configuration of data points and the local rate of convergence. Here, we limit ourselves to $p = 2$. In this section, we see how (because of the curvature) the speed of convergence in locating the center of mass depends on the configuration of the data points (this phenomenon is not present in a Euclidean space, see below). In particular, we give a (partial) qualitative answer to the following question:"For which configurations of data points does Algorithm 1 converge very fast? very slowly?"

From Theorem 2.12 and the definition of q in (2.14) it is clear that in addition to α, the ratio $\frac{h_S}{H_S}$ is also important in determining the speed of convergence, and the asymptotic speed of convergence depends on the ratio $\frac{h_S}{H_S}$ in a very small neighborhood S around $\bar{x}$. In the Euclidean case the ratio is 1 and with $\alpha = 1$ we have $q = 0$; therefore, Algorithm 1 finds the center of mass in *one* step! See (2.13) and (2.14). However, in a curved manifold the ratio $\frac{h_S}{H_S}$ can be very small, due to drastic difference in the behavior of the Hessian of the distance function along different directions. Next we give simple examples that demonstrate this fact.

We consider the case of constant curvature, since in this case the eigenvalues of the Hessian of the distance function are the same along all directions but the radial direction. Furthermore, let us assume M is a 2-dimensional simply connected manifolds with constant curvature, that is $M = \mathbb{S}^2_\Delta$ where $\Delta = 1$ or $\Delta = -1$ (with the convention $\mathbb{S}^2_1 \equiv \mathbb{S}^2$). We construct two simple configurations for which Algorithm 1 converges very fast and very slowly, respectively. Consider four data points $\{x_i\}_{i=1}^4$ and the closed ball $\bar{B}(o,\rho) \subset M$, where $\rho < r_{\mathrm{cx}}$. Assume x_1 and x_2 are on the boundary of the ball in antipodal positions and that x_3 and x_4 are also in antipodal positions such that the geodesic γ_{ox_1} from o to x_1 and the geodesic γ_{o,x_3} from o to x_3 are perpendicular at o. We denote this configuration by ⁘. Obviously, $\bar{x} = o$ is the center of mass of $\{x_i\}_{i=1}^4$ with equal weights. It is easy to verify that for the ⁘ configuration the Hessian of f_2 at $x = o$ both along γ_{ox_1} and along γ_{ox_3} has eigenvalue $\frac{1}{2}(\rho\,\mathrm{ct}_\Delta(\rho) + 1)$. Consequently, at o the ratio of the smallest and largest eigenvalue is 1, hence $\frac{h_S}{H_S} \approx 1$ around $\bar{x} = o$; and therefore, one expects that the local rate of convergence will be very fast. The opposite configuration is ⁚, that is, when x_3 and x_4 coincide with x_1 and x_2, respectively. In this case, at $x = o$ along γ_{ox_1} the

Hessian of f_2 has eigenvalue of 1 and in the perpendicular direction it has eigenvalue $\rho\mathrm{ct}_\Delta(\rho)$. Therefore, if $\Delta = 1$, we have $\frac{h_S}{H_S} \approx \rho\cot\rho$ around o which, in particular, can be very small if ρ is close to $\frac{\pi}{2}$. If $\Delta = -1$, we have $\frac{h_S}{H_S} \approx (\rho\coth\rho)^{-1}$, which again can be small if ρ is large. It is well known that the shape of the level sets of a function in a neighborhood of a minimizer is related to the ratio $\frac{h_S}{H_S}$. If the level sets are very elongated or thin, this means that the Hessian has very small eigenvalues along longitudinal directions and very large eigenvalues along the lateral directions and hence $\frac{h_S}{H_S}$ can be very small. For our two configurations, we encourage the reader to compare the shapes of the level sets of f_2 in $B(o,\rho) \subset \mathbb{S}^2$ for levels close to $f_2(o)$ (especially when ρ is close to $\frac{\pi}{2}$) with the level sets of f_2 in $B(o,\rho) \subset \mathbb{S}^2_{-1}$ for levels close to $f_2(o)$ when ρ is very large).

As a tangible example, on the standard unit sphere $\mathbb{S}^2$ we run Algorithm 1 for both the configurations with two different values of $\rho \approx 0.35\pi$ and $\rho \approx 0.47\pi$. The initial condition is chosen *randomly*. The step-size is chosen as $t_k = 1$. Figure 5.1 shows the distance $d(x^k, \bar{x})$ in terms of the iteration index k. It is clear that for the ⁚ configuration the convergence is slower than the convergence for the •⁚• configuration, and as ρ increases, convergence for both configurations becomes slower. However, for the ⁚ configuration as ρ approaches $\frac{\pi}{2}$, the convergence becomes extremely slow and the •⁚• configuration is much more robust in that sense. Note that when $\rho \approx \frac{\pi}{2}$ even the center of mass of the ⁚ configuration is on the verge of non-uniqueness and this causes further (error) sensitivity and hence poor convergence (see [2] on the issue of high noise-sensitivity of the Riemannian mean in positively curved manifolds).

Although our example is rare in statistical applications, in a more general setting also one expects that if the configuration of data points is such that the convex hull of the data points has an elongated shape (especially if the length of the convex hull is large), then locating the Riemannian center of mass becomes a difficult problem (with the exception of the Euclidean case). Our analysis does not tell the whole story in the case of variable curvature and we need more detailed analysis that takes into account the variability of eigenvalues of the Hessian of the distance function along non-radial directions, as well.

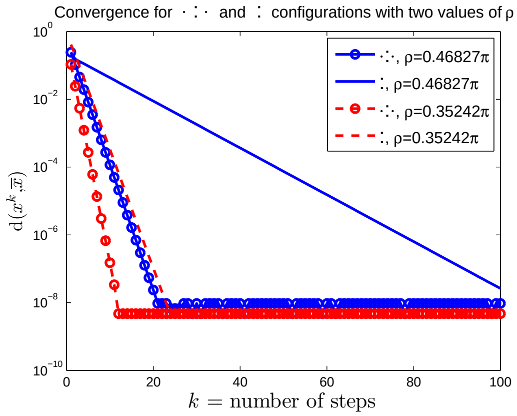

Fig. 5.1: Convergence behavior of Algorithm 1 for locating the center of mass two configurations denoted by •⁚• and ⁚ on the unit sphere $\mathbb{S}^2$. The step-size is $t_k = 1$ and the initial condition is chosen randomly.

6. Global Convergence on $SO(3)$, $\mathbb{S}^n$ and Similar Manifolds. So far, our focus was on finding the global Riemannian center of mass using a constant step-size gradient descent algorithm; hence, we only studied the *local behavior* of the algorithm (albeit in relatively large domains). Studying the *global behavior* of the algorithm (i.e., convergence for arbitrary initial conditions) is more subtle. Finding the global center of mass by gradient descent under arbitrary initial condition is out of question (in this case or a more general one where the data points are not localized one could you stochastic global optimization methods e.g., [6] or (semi-) combinatorial methods e.g., [27, 13]). Even convergence to a local center is not straightforward due to the fact that f_p is not differentiable globally. One can verify that f_p is a locally Lipschitz function and hence differentiable almost everywhere. In this section we want to see to what extent simple constant step-size gradient descent (with as little modification as possible) still could work and find a center of mass (local or global). Our approach is partly based on our work [45], in which similar situations in a different application were addressed. Our discussion is limited to $p = 2$.

6.1. Riemannian center of mass and the cut loci of data points. Under certain conditions on the manifold M, it is easy to guarantee that the crucial relation (2.10) holds. One example of such conditions is Condition (L):

> Condition (L): M is compact and for every point $y \in M$ every cut point of y is a local maximizer of the distance function $x \mapsto d(x, y)$.

The reader can check that $\mathbb{S}^n$, $SO(3)$, the real projective plane $\mathbb{RP}^n$, and the n-torus $\mathbb{T}^n$ with their standard Riemannian metrics satisfy condition (L).

Remark 6.1. Recall that by the standard Riemannian metric on $SO(3)$ we mean the bi-invariant metric which at I_3, the identity of $SO(3)$, is defined as $\langle X, Y\rangle_{I_3} = \frac{1}{2}\text{trace}(X^\top Y)$, where X, Y are (3×3 skew-symmetric) tangent vectors at I_3. In this metric the distance between x and y in $SO(3)$ is $d(x, y) = \frac{1}{\sqrt{2}}\|\log x^\top y\|_F$, where $\log(\cdot)$ denotes the matrix logarithm (if $x^\top y$ has eigenvalues of -1, then $d(x, y) = \pi$ and x and y are cut points of each other). Moreover, in this metric inj$SO(3) = \pi$ and $\delta = \Delta = \frac{1}{4}$. Similarly, in the standard metric for the (unit radius) real projective plane $\mathbb{RP}^n$ ($n \geq 2$), we have $\delta = \Delta = 1$ and inj$\mathbb{RP}^n = \frac{\pi}{2}$ (for $n = 1$ obviously $\delta = \Delta = 0$). Also it is known that in its standard metric, $SO(3)$ is isometric to the real projective plane $\mathbb{RP}^3_{\frac{1}{4}}$, which is the real projective plane in $\mathbb{R}^4$ with radius 2.

The main implication of Condition (L) is that the distance to y remains constant in $\mathcal{C}_y$ and that, in fact, a cut point of y is a global maximizer of the distance function, as the next proposition shows. The proof of the proposition requires some deep results from Riemannian geometry, but it is easy to explicitly verify the proposition in above manifolds. Most likely further detailed characterization of manifolds satisfying Condition (L) is possible (or may exist in the literature), but that is beyond the scope of this paper.

Proposition 6.2. *Let M satisfy Condition (L). Then for every $y \in M$ the function $x \mapsto d(x, y)$ is constant on $\mathcal{C}_y$. Therefore, any cut point of y is a global maximizer of $x \mapsto d(x, y)$. Moreover, there are at least two minimal geodesics from y to $x \in \mathcal{C}_y$, and for any such geodesic there is another (mirror) one such that the two fit smoothly together to form a closed geodesic.*

Proof. It is known that for a compact manifold $\mathcal{C}_y$ is a connected set (see e.g., [15, p. 95] or [42, p. 208]). Therefore, since every $x \in \mathcal{C}_y$ is local maximizer of $x \mapsto d(x, y)$, $d(x, y)$ must remain constant on $\mathcal{C}_x$. In fact, we have $d(x, y) = $ inj$y$ on $\mathcal{C}_y$. It is also known that if x is a local maximizer of $x \mapsto d(x, y)$, then it is a critical point (see [40, p. 355] for exact definition of a critical point, which is different from the standard definition in calculus), this implies that there are at least two minimal geodesics from y to x (i.e., x is an ordinary cut point). Then it follows from a result due to Klingenberg (see e.g. [40, p. 142, Lemma 16]) that for any such geodesic there is another (mirror) one such that they fit each other smoothly to form a closed geodesic. □

At a point $x \in \mathcal{C}_{x_i} \subset M$ we can group the terms in f_2 in two groups: one which is comprised of functions smooth at x denoted by f_2^{s} and one which is not smooth at x denoted by f_2^{ns} (this term simply is the term containing $d^2(x, x_i)$); and we write $f_2 = f_2^{\mathrm{ns}} + f_2^{\mathrm{s}}$. We call f_2^{s} the *smooth part* of f_2 and f_2^{ns} the *nonsmooth part* of f_2 at x. In general, x can belong to more than one cut locus, but in that case also this decomposition remains valid.

The following result shows that a cut point of a data point cannot be a local (or global) center of mass in manifolds satisfying condition (L). It is an extension of results in [13, 27].

THEOREM 6.3. *If M satisfies condition (L), then no local Riemannian center of mass of $\{x_i\}_{i=1}^N$ belongs to the cut loci of the data points. In particular, any local Riemannian center of mass of $\{x_i\}_{i=1}^N$ is a zero of the gradient of f_2.*

Proof. Let $\bar{x}$ be a local minimizer of f_2 which belongs to the cut loci of the data points. First, let us assume that $\bar{x}$ belongs to the cut locus of exactly one of the data points, say x_1. We can write $f_2 = f_2^{\mathrm{ns}} + f_2^{\mathrm{s}}$, where f_2^{s} is smooth at $\bar{x}$ and f_2^{ns} is non-differentiable at $\bar{x}$ but has directional derivatives and has a maximizer at $\bar{x}$. Notice that f_2^{ns} has negative directional derivative in some directions leaving $\bar{x}$ (e.g., along the two directions to x_1). Now $\bar{x}$ being a local minimizer of f_2 requires the gradient of the smooth part f_2^{s} be zero, but that is not enough to balance the negative directional derivatives of the non-smooth part f_2^{ns}, hence, $\bar{x}$ cannot be a local minimizer of f_2. The same argument applies if $\bar{x}$ belongs to more than one cut locus. □

6.2. Almost gradients. If x belongs to the cut locus of x_i, $\mathcal{C}_{x_i}$, then according to the standard definition, $\exp_x^{-1} x_i$ is not defined. However, under Condition (L) we can define $\exp_x^{-1} x_i$ in such a way that a well-defined algorithm results. By Proposition 6.2, there are more than one minimal geodesics from x to x_i. Let $\gamma_{xx_i} : [0,1] \to M$ be one such geodesic (note that $\gamma_{xx_i}(0) = x$ and $\gamma_{xx_i}(1) = x_i$). We call the negative of the initial velocity of this geodesic, namely $-\dot{\gamma}_{xx_i}(0^+)$, an *almost gradient* of $z \mapsto \frac{1}{2}d^2(z, x_i)$ at $z = x \in \mathcal{C}_{x_i}$.[16] Recall that $z \mapsto \frac{1}{2}d^2(z, x_i)$ is smooth in $M/\mathcal{C}_{x_i}$, and notice that an almost gradient at $x \in \mathcal{C}_{x_i}$ is, in fact, the limit of a sequence of gradients of $z \mapsto \frac{1}{2}d^2(z, x_i)$ evaluated at $\langle x^k \rangle_k$, where $\langle x^k \rangle_k$ is a sequence of points in $M/\mathcal{C}_{x_i}$ converging to $x \in \mathcal{C}_{x_i}$. Obviously, by replacing $\exp_x^{-1} x_i$ in (2.8) with $-\dot{\gamma}_{xx_i}(0^+)$ we can define an almost gradient for f_2 at $x \in \mathcal{C}_{x_i}$.

Due to Condition (L) there is a mirror geodesic to γ_{xx_i}, namely $\tilde{\gamma}_{xx_i} : [0,1] \to M$, where $\tilde{\gamma}_{xx_i}(0) = x$, $\tilde{\gamma}_{xx_i}(1) = x_i$, and $\dot{\tilde{\gamma}}_{xx_i}(0^+) = -\dot{\gamma}_{xx_i}(0^+)$. There is a possibility that with the choice $\exp_x^{-1} x_i = -\dot{\gamma}_{xx_i}(0^+)$ in (2.8), $\nabla f_2(x)$ becomes zero (note that

[16] Generalizations of the notion of gradient have appeared in the literature on nonsmooth optimization under various names and forms. Our definition is essentially the same as the almost gradient introduced by Shor [44]. The notion of almost gradient differs from that of *generalized gradient* [17], which is the convex hull of almost gradients. For our current application, we find almost gradients more convenient than generalized gradients. In Riemannian geometry literature also the term "generalized gradient" is often used to denote a set of vectors and not a single vector [40, Ch. 11]. It is interesting to mention that in Riemannian geometry, the powerful theory of critical points of distance functions, which has brought about some profound results, is essentially the study of the generalized gradient of the Riemannian distance function on a manifold and its relation to the topology of the manifold (see [40, Ch. 11]).

this does not contradict Theorem 6.3). As an example, in Example 2.11 with $w_1 = \frac{3}{8}$, this situation happens at $x = x_1'$ for the clockwise geodesic from x_1' to x_1 (see Figure 2.1). In such a case, we can simply choose $\exp_x^{-1} x_i = -\dot{\tilde{\gamma}}_{xx_i}(0)$, which obviously results in non-zero $\nabla f_2(x)$. Notice that $\nabla f_2(x)$, defined in this fashion, is clearly a descent direction at x. If x belongs to more than one cut locus, our definition extends similarly. We call $\nabla f_2(x)$, defined in this fashion, a *preferred almost gradient* of f_2 at x. In the sequel, unless mentioned explicitly, by ∇f_2 at a cut point we mean a preferred almost gradient (i.e., one that is not zero), and we implement Algorithm 1 with such ∇f_2. Therefore, we have defined descent directions on the entire M. However, note that the caveat is that ∇f_2 is not continuous at a cut point and that is a major obstacle in our algorithm having a completely desirable behavior. Also note that the discontinuity is the result of the inherent non-uniqueness of almost gradients and not our preferred choice. We could implement Algorithm 1 with any almost gradient, but there is a chance that the algorithm would stop at a cut point of a data point in finite steps (although in a generic case the chance of this happening is zero). However, with the preferred almost gradient this possibility is removed, which is conceptually desirable.

Now, we verify that (2.10) remains valid under Condition (L). First, we show that the $s \mapsto \exp_x(-s\nabla f(x))$ meets any cut locus $\mathcal{C}_{x_i}$ in a well-behaved manner.

Proposition 6.4. *Let M satisfy Condition (L) and define $c(s) = \exp_x(-s\nabla f(x))$ where $s \in [0,t]$. The set $I = \{s \in (0,t) | c(s) \notin \mathcal{C}_{x_i}\}$ is a union of (at most) countable disjoint open subintervals of $[0,t]$.*

Proof. Consider the continuous function $z(s) = d(x_i, c(s))$. Due to condition (L) we have $c(s) \notin \mathcal{C}_{x_i}$ if and only if $z(s) < \mathrm{inj} x_i$. This implies that I is an open set in $(0,t)$. The claim follows from a well-known result about open sets in $\mathbb{R}$. □

Proposition 6.5. *Let M satisfy Condition (L) and $\{x_i\}_{i=1}^N \subset M$ be a given set of data points (not necessarily localized in a small region). Let H_M be an upper bound on the eigenvalues of the Hessian of $x \mapsto \frac{1}{2}d^2(x, x_i)$ (wherever defined) for every i. (H_M is automatically an upper bound on the eigenvalues of the Hessian of f_2 (wherever defined) and if $\Delta \geq 0$ then we can take $H_M = 1$.) In Algorithm 1 with the preferred almost gradient at cut loci of the data points choose $t \in (0, \frac{2}{H_M})$. Then for f_2 relation (2.10) holds (with $H_S = H_M$), with equality only if x is a zero of the gradient of f_2.*

Proof. For convenience we replace the role of t in (2.10) with $s \in [0,t]$. Set $c(s) = \exp_x(-s\nabla f_2(x))$ and denote the left hand and right hand sides of (2.10) by $f_2(s;x)$ and $\tilde{f}_2(s;x)$, respectively (note that $f_2(s;x) = f_2(c(s))$). For now assume that x does not belong to any cut loci of the data points. Assume that $c(s)$ meets at least one cut locus, otherwise (2.10) holds trivially. For now assume that $c(s)$ only meets C_{x_1}. In particular, (by Proposition 6.4) let s_1 be the first point such that $c(s_1) \in \mathcal{C}_{x_1}$. In $(0, s_1)$, (2.10) holds true and $f_2(s;x)$ is C^2. Note that by Proposition 6.4 either $c(s)$ leaves $\mathcal{C}_{x_1}$ immediately, that is, $c((s_1, s_2)) \notin C_{x_1}$ for a (maximal) $s_2 > s_1$ or it stays in $\mathcal{C}_{x_1}$, that is, $c([s_1, s_2]) \in C_{x_1}$ for a (maximal) $s_2 > s_1$. A relation not explicit in Proposition 2.7 is between the derivatives of f_2 and $\tilde{f}_2$, namely that $f_2'(s;x) \leq \tilde{f}_2'(s;x)$ for $s \in (0, s_1)$. Having this in mind, in the first mentioned case we have $f_2(s_1^-;x) \leq \tilde{f}_2(s_1^-;x)$, $f_2'(s_1^+;x) \leq f_2'(s_1^-;x) \leq \tilde{f}_2'(s_1^-;x) = \tilde{f}_2'(s_1^+;x)$ (because of Condition (L)), and $f_2''(s_1^+;x) \leq H_M$. Therefore clearly the relation $f_2(s;x) \leq \tilde{f}_2(s;x)$ holds for $s \in (s_1, s_2)$. In the second case, the nonsmooth part of $f_2(s;x)$ remains constant in $\mathcal{C}_{x_1}$ so we clearly have $f_2^{'\mathrm{ns}}(s_1^-;x) > f_2^{'\mathrm{ns}}(s_1^+;x) = 0$ and hence again $f_2'(s_1^+;x) \leq f_2'(s_1^-;x) \leq \tilde{f}_2'(s_1^-;x) = \tilde{f}_2'(s_1^+;x)$. Moreover, for $s \in (s_1, s_2)$

we can assume that $f_2^{\prime \mathrm{ns}}(s;x)$ is (constant) smooth, hence $f_2(s;x)$ can be assumed smooth in (s_1, s_2) with $f_2''(s;x) \leq H_M$. This again implies $f_2(s_2;x) \leq \tilde{f}_2(s_2;x)$ and $f_2'(s_2^-;x) \leq \tilde{f}_2(s_2;x)$. Because of Condition (L) we must have $f_2'(s_2^+;x) \leq f_2'(s_2^-;x)$ and hence $f_2(s;x) \leq \tilde{f}_2(s;x)$ for $s \in (s_2, s_3)$ where $s_3 \leq t$ is the next time $c(s)$ enters $\mathcal{C}_{x_1}$. This argument can be repeated if $c(s)$ meets $\mathcal{C}_{x_1}$ more. Moreover, the same argument can be extended to the case where $c(s)$ meets more than one cut locus or if x belongs to a cut locus (in which a preferred almost is employed). □

6.3. Global convergence. Now we prove our global convergence result, which states that any accumulation point of the algorithm is either a zero of ∇f_2 (at which f_2 is smooth) or a cut point at which an almost gradient is zero. This latter scenario is a rather peculiar (and rare) scenario stemming from discontinuity of ∇f_2 at cut points. In the proof, we use the fact that if $x \in M$ does not belong to any of the cut loci, then for small r, $B(x, r)$ also does not intersect any of the cut loci.

Theorem 6.6. *Assume $p = 2$ and let M satisfy Condition (L) and let $\{x_i\}_{i=1}^N \subset M$ be a set of data points (not necessarily localized in a small region). Let $t_k = t \in (0, \frac{2}{H_M})$ be the step-size. Then any accumulation point of Algorithm 1 implemented with the preferred almost gradient is either a zero of the gradient of f_2 (in particular f_2 is smooth at such a point) or a cut point of one of the data points at which at least one almost gradient of f_2 is zero. However, such a cut point is neither a local minimizer of f_2 nor a fixed point of the algorithm, hence it is an unstable accumulation point, in the sense that with small random noise added to the iterates, that point will not be an accumulation point anymore.*

Proof. Since relation (2.10) holds, the proof is essentially the same as the first part of the proof of Theorem 2.10, except that from $\|\nabla f_2(x^{k_j})\| \to 0$ we conclude that either x^* is a zero of the gradient of f_2 or it is a cut point at which an almost gradient is zero. To see this note that if x^* is not a cut point of any data point, then ∇f_2 is continuous in a neighborhood of x^*, hence we must have $\nabla f_2(x^*) = 0$. If x^* is a cut point, then $\|\nabla f_2(x^{k_j})\| \to 0$ exactly means that an almost gradient at x^* is zero. The statement about instability is obvious. □

For an almost gradient to be zero at a given cut point there must be a relation between the data points and the weights w_i's (we see this in the above example in relation to Example 2.11). Therefore, generically (i.e., arbitrary data points and arbitrary weights), the probability of an almost gradient at a cut point being zero is diminished. Hence, generically, one expects that all accumulations points are zeros of the gradient of f_2.

Corollary 6.7. *Let M be either $SO(3)$, $\mathbb{RP}^n$, or $\mathbb{T}^n$. Furthermore, assume that no cut point of the data points has a zero almost gradient. Then the iterates converge to a local Riemannian center of mass.*

Proof. In view of Theorem 6.6 and Proposition 2.14, it suffices to show that any zero of ∇f_2 is a non-degenerate local minimizer. By Theorem 6.3, a local minimizer of f_2 is not a cut locus. Let $\bar{x}$ be an accumulation point, clearly $\bar{x}$ must be zero of gradient ∇f_2 and not a cut point of any of data points. This means that $\{x_i\}_{i=1}^N \subset B(\bar{x}, \rho)$, where for $SO(3)$, $\mathbb{RP}^n$, and $\mathbb{T}^n$ we have $\rho < \pi$, $\rho < \frac{\pi}{2}$, $\rho < \pi$, respectively. Notice that f_2 is C^2 in $B(\bar{x}, \epsilon)$ for small ϵ. It follows from (2.5) that, in each case, f_2 is strictly convex in $B(\bar{x}, \epsilon)$, hence $\bar{x}$ must be a non-degenerate local minimizer. (In fact, there is $h_\rho > 0$ for which the eigenvalues of the Hessian of f_2 at $\bar{x}$ are not smaller than h_ρ: For $SO(3)$, $\mathbb{RP}^n (n > 2)$, and $\mathbb{T}^n$ we have $h_\rho = \frac{\rho}{2} \cot \frac{\rho}{2}, \rho \cot \rho$, and 1, respectively. For $\mathbb{RP}^1$ we have $h_\rho = 1$.) Since $\bar{x}$ is an accumulation point, by Proposition 2.14 (or similar results), this is enough to guarantee that the iterates converge to $\bar{x}$. □

Acknowledgements. The authors thank the anonymous referees for their diligent review and insightful comments. Several corrections and improvements over the initial version of the paper are due to the referees' comments and suggestions. Authors thank Marc Arnaudon for allowing us to include his proof of Theorem 3.1. BA is grateful to David Groisser and Huiling Le for insightful discussions and useful suggestions.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, Princeton, NJ, 2008.

[2] B. Afsari. *Means and Averaging on Riemannian Manifolds.* PhD thesis, University of Maryland, College Park, Dec. 2009.

[3] B. Afsari. Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proc. Amer. Math. Soc.*, 139:655–673, 2011.

[4] M. Arnaudon. Private communication, May 2012.

[5] M. Arnaudon, C. Dombry, A. Phan, and L. Yang. Stochastic algorithms for computing means of probability measures. *Stochastic Processes and their Applications*, 122(4):1437–1455, April 2012.

[6] M. Arnaudon and L. Miclo. Means in complete manifolds: uniqueness and approximation. arXiv:1207.3232v1, July 2012.

[7] M. Arnaudon and F. Nielsen. On approximating the Riemannian 1-center. *Computational Geometry*, 46(1):93–104, January 2013.

[8] M. Berger. *A Panoramic View of Riemannian Geometry.* Springer, 2007.

[9] A. Bhattacharya. *Nonparametric Statistics on Manifolds with Applications to Shape Spaces.* PhD thesis, The University of Arizona, 2008.

[10] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.*, 31(1):1–29, 2003.

[11] R. L. Bishop. Decomposition of cut loci. *Proceedings of the American Mathematical Society*, 65(1):133–136, July 1977.

[12] S. R. Buss and J. P. Fillmore. Spherical averages and application to spherical splines and interpolation. *ACM Transcations on Graphics*, 20(2):95–126, April 2001.

[13] B. Charlier. Necessery and sufficient condition for the existence of a frchet mean on the circle. *ESAIM: Probability and Statistics*, eFirst, 2012.

[14] I. Chavel. *Riemannian Geometry: A Modern Introduction.* Cambridge University Press, 2nd, 2006.

[15] J. Cheeger and D. Ebin. *Comparison theorems in Riemannian geometry.* AMS Chelsea Publishing, Providence RI, 2008.

[16] G. S. Chirikjian. *Stochastic Models, Information Theory, Analytic Methods and Modern Applications and Lie Groups*, volume 2. Birkhäuser, 2011.

[17] F. H. Clarke. *Optimization and Nonsmooth Analysis.* SIAM, 1990.

[18] S. Fiori. Solving minimal-distance problems over the manifold of real symplectic matrices. *SIAM Journal on Matrix Analysis and Applications*, 32(3):938–968, 2011.

[19] S. Fiori and T. Tanaka. An algorithm to compute averages on matrix Lie groups. *IEEE Transactions on Signal Processing*, 57(12):4734–4743, December 2009.

[20] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, March 2009.

[21] A. Goh, C. Lenglet, P. Thompson, and R. Vidal. A nonparametric Riemannian framework for processing high angular resolution diffusion images (HARDI). In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[22] U. Grenander. *Probabilities on Algebraic Structures.* John Wiley and Sons (reprinted by Dover, 2008), 1963.

[23] D. Groisser. Newton's method, zeros of vector fields, and the Riemannian center of mass. *Adv. in Appl. Math.*, 33:95–135, Nov 2004.

[24] D. Groisser. On the convergence of some Procrustean averaging algorithms. *Stochastics*, 77(1):31–60, February 2005.

[25] K. Grove and H. Karcher. How to conjugate C^1-close group actions? *Math. Z.*, 132(1):11–20, March 1973.

[26] R. Hartley, K. Aftab, and J. Trumpf. L^1 rotation averaging using the Weiszfeld algorithm. In

IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[27] T. Hotz and S. Huckemann. Intrinsic means on the circle: Uniqueness, locus and asymptotics. arXiv:1108.2141v1, 2011.

[28] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30:509–541, September 1977.

[29] D. G. Kendall, D. Barden, T. K. Carne, and H. Le. *Shape and Shape Theory.* Wiley Series In Probability And Statistics. John Wiley & Sons, 1999.

[30] W. S. Kendall. Probability, convexity, and harmonic maps with small image I: Uniqueness and fine existence. *Proc. Lond. Math. Soc.*, 61(2):371–406, 2 1990.

[31] K. A. Krakowski. *Geometrical Methods of Inference.* PhD thesis, The University of Western Australia, Aug. 2002.

[32] K. A. Krakowski, K. Hüper, and J. H. Manton. On the computation of the Karcher mean on spheres and special orthogonal groups. In H. Araújo and M. I. Ribeiro, editors, *Workshop on Robotics and Mathematics (ROBOMAT 07)*, pages 119–124, Coimbra, Portugal, September 2007.

[33] H. Le. Locating Fréchet means with application to shape spaces. *Adv. in Appl. Probab.*, 33(2):324–338, July 2001.

[34] H. Le. Estimation of Reimannian barycenters. *LMS Journal of Computation and Mathematics*, 7:193–200, 2004.

[35] J. H. Manton. A globally convergent numerical algorithm for computing the centre of mass on compact Lie groups. In *Proceedings of the Eighth International Conference on Control, Automation, Robotics and Vision*, pages 2211–2216, Kunming, China, December 2004.

[36] W. J. M'Clellanp and T. Preston. *A Treatise On Spherical Trigonometry, with numerous examples Part I.* Macmillan And Co., 1886. Available at Google Books.

[37] M. Moakher. Means and averaging in the group of rotations. *SIAM Journal on Matrix Analysis and Applications*, 24(1):1–16, 2002.

[38] X. Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vision*, 25(1):127–154, July 2006.

[39] X. Pennec and V. Arsigny. *Matrix Information Geometry*, chapter Exponential Barycenters of the Canonical Cartan Connection and Invariant Means on Lie Groups. Springer, 2012.

[40] P. Petersen. *Riemannian Geometry.* Springer, 2006.

[41] B. Polyak. *Introduction to Optimization.* Translations Series in Mathematics and Engineering. Optimization Software, 1987.

[42] T. Sakai. *Riemannian Geometry*, volume 149. American Mathematical Society, 1996.

[43] A. Sarlette and R. Sepulchre. Consensus optimization on manifolds. *SIAM Journal of Control and Optimization*, 2008. to appear.

[44] N. Z. Shor. *Minimization Methods for Non-Differentiable Functions*, volume 3 of *Springer Series in Computational Mathematics.* Springer-Verlag, 1985.

[45] R. Tron, B. Afsari, and R. Vidal. Intrinsic consensus on $SO(3)$ with almost-global convergence. In *51st Annual IEEE Conference on Decision and Control (CDC)*, pages 2052–2058, 2012.

[46] R. Tron, R. Vidal, and A. Terzis. Distributed pose averaging in camera networks via consensus on $SE(3)$. In *International Conference on Distributed Smart Cameras*, 2008.

[47] P. Turaga, A. Veeraraghavan, and R. Chellappa. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[48] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, October 2008.

[49] C. Udrişte. *Convex Functions and Optimization Methods on Riemannian Manifolds.* Mathematics and Its Applications. Kluwer Academic Publishers, 1994.

[50] L. Yang. Riemannian median and its estimation. *LMS Journal of Computations and Mathematics*, 13:461–479, 2010.

[51] S-T. Yau. Non-existence of continuous convex functions on certain Riemannian manifolds. *Mathematische Annalen*, 207:269–270, 1974.